



The Sheku Bayoh Public Inquiry

Witness Statement

Dr Nicholas Schurch

Taken by [REDACTED] on MS Teams on Monday 07 November 2022

Witness details and Professional background

1. My name is Dr Nicholas [REDACTED] Schurch. I was born in 1977. My contact details are known to the Inquiry.
2. I am a Quantitative and Senior Research scientist with over 20 years of experience in Statistics, Modern Data Analysis and Data Visualisation Techniques spanning a wide range of quantitative science areas.
3. My current role is Principal Statistician for Environmental Science and Ecology with Biomathematics and Statistics Scotland (BioSS), and I have been in this post since 2019. Prior to this role, I spent ten years working as a Bioinformatician and Data Scientist at the University of Dundee. Before that, I spent approximately ten years working as an Astrophysicist in a number of roles, following my PHD in Astrophysics and my undergraduate degree in Physics with Astrophysics. The full details of my experience are set out in my CV which forms part of my statistical report.


Statistical Report – Section 2.1-2.2 Introduction and Data

4. I have been asked why I was instructed by Professor Dawson to carry out the Soil Elemental Composition Statistical Similarity Assessment. Professor Dawson instructed me to conduct a statistical analysis of the data originally generated as part of her initial analysis. This data comes from a technique called 'Energy Dispersive X-Ray Analysis', which is used to analyse soil samples. It generates a type of data that statisticians refer to as 'relative composition data'. In this instance, a straightforward way of understanding this concept is that each

Signature of witness..... [REDACTED]

measurement is a set of percentages, and those percentages always add up to 100%. This type of data requires expertise to work with effectively because many standard statistical tools/approaches that scientists use make assumptions that are violated by this kind of data. Professor Dawson looked at this data originally and was confronted by some of these problems and decided that she required additional expertise to analyse the data appropriately.

5. The purpose of the statistical analysis was to assess the similarity of a set of questioned soil samples with several sets of reference samples and to try to conclude whether the questioned samples were consistent with sharing a common origin with any or all of those reference soil samples. The reference soil samples in this case are the soil samples taken from the known locations (the boots of Sheku Bayoh and PC Craig Walker), as opposed to the questioned sample (the vest of Nicole Short).
6. I have been asked why standard statistical approaches are unsuitable for use with relative compositional data. Many standard statistical approaches are unsuitable for use with relative compositional data, because they commonly assume that the data are independent and normally distributed. Neither of those is the case for relative compositional data as a normal distribution doesn't have a start or end point. It can extend to both sides of the number range, whereas percentages are obviously limited; you cannot go below 0% and you cannot go above 100%. If you change one of the percentages in a relative composition sample, you change one or more of the other percentages. For instance, if you had something that was 60%-40% percentage and then you increase the 60% to 65%, you would have to decrease the 40% to 35% as those two measurements are not independent of each other. A straightforward way of understanding this would be in the context of the Brexit vote, which was 52% (for) - 48% (against). If the 52% goes up to 55%, the 48% goes down to 45%, and so the measurements of 52% for and 48% against are not independent of each other. They are related. This presents difficulties and as such I have taken approaches to deal with those properties to allow us to use statistical approaches such as Clustering and Principal Component Analysis.
7. Now, when there are two categories (such as above with Brexit), this sort of interdependence is obvious. However, if you imagine having a set of ten different percentages, that all add up to 100%, if you change one of those percentages it is unclear which of the other percentages will change. It might be all of them, it might be only one of them or it might be a subset of them. It becomes increasingly challenging, and a lot less intuitive, to work with these data when there are more categories, as the question becomes, what will change and what does this change mean?
8. In this case, we had eleven measurements/categories which are the eleven elements that form the composition of each soil sample. So, imagine comparing two of those samples. If you see a difference in one of those eleven categories and you want to know whether that difference is meaningful, not only do you have

Signature of witness..... 

to look at how that change has affected the other categories, but you need to look at the pattern across all of the categories. It is not straightforward to intuitively understand how changes have knock-on effects. This is why we try to transform the data to a form where the values are independent of each other, and we do this by forming what we call log ratios.

9. For transforming the data into log ratios, we identify one of the eleven categories as a reference category. Ideally, this reference category should have a some convenient characteristics: it should be present in all of the different samples, it should be relatively consistent across all of the samples, and it should not show too much variability. According to Professor Dawson, aluminium is common to most soils and would make a good reference category. According to Professor Dawson, it is present at reasonable levels in most soils, and it also tends to not vary strongly between soil types. After identifying the reference category, you then take the ratio of each of the other categories (for example magnesium and silicon) to that category, and that produces transformed values that are independent of each other. A data point refers to a single measurement of elemental percentage within a sample. To stress further, we have the soil samples, each of which has eleven data points but we select one of those data points as a reference sample (aluminium) in order to produce ratios; so we end up with ten ratio data values for each sample. Importantly, these ratios are now independent of each other even if the compositional data points underneath are not.
10. These ratios are still not ideal because they cannot be less than zero. Furthermore, ratios less than one are compressed into the range zero to one, whereas ratios greater than one have no upper limit and can “spread out” a lot more. To resolve this asymmetry we take logarithms of the ratios with the result that equal space is given to the ratios less than one and the ratios that are greater than one. As a result, we end up a suite of log ratios that are unbounded, are approximately normally distributed, and are independent. A logarithm is a standard mathematical transform. It is important to note that there are rare instances where the underlying percentage data points have unusual properties such that this process produces results that are not perfectly normally distributed.
11. I have been asked what a ‘replicate’ is and why the replication of a sample is important. A replicate is simply a repeated measurement, usually taken at the same time and/or from the same location. However, there is some nuance to the term because replication can be done at several distinct levels. For example, if you have a sample and you take several measurements of the same property (e.g., weight, acidity, size) on that sample, this is what is commonly referred to as ‘technical replication’, because you are in effect assessing the accuracy or the repeatability of a measurement process. As a simple analogy, imagine measuring your height with a tape measure. You have measured it once and you have a measurement. You think that you have measured your height, but you do not know how accurate the tape measure is, and perhaps you slouched or maybe you slumped, and as a result you might want to measure your height several

Signature of witness..... 

different times with the same tape measure. The tape measure will have an accuracy as well, so these will be 'technical replications'. These are measurements of the same sample, with ideally the same tools, taken multiple times and this characterises the variability that you see in your experimental setup/tools.

12. As it relates to the soil samples, what we have done here is to take repeat samples from a larger quantity. Imagine having a large piece of soil and you take several smaller pieces from within that soil and conduct the same analysis on each of those pieces. The variability between the results of those different sub-samples, those 'replicates', is affected not only by your experimental setup but is also affected by the variation within the larger sample. A good analogy for this would be if you wanted to measure the height of children. You might measure the height of one child six times with a tape measure and then calculate an average, to hopefully obtain a more accurate height for that one child. You will also obtain an idea of how accurate your tape measure is and how much variability there is in that measurement. However, you might also measure ten children out of a class of thirty to get an idea of what the height variation across that class is. In this instance, we have soil samples, and we have soil material that is taken from a particular location, (for example the two pairs of boots and the vest). The small pieces of that soil material from each of the soil samples is taken for analysis and we have six replicates within each soil sample.
13. Understanding the variation within the samples is an important consideration for statisticians. We are trying to understand the variation in quantities that we measure in lots of different areas of science. For example, political scientists want to understand variation in voting patterns. Essentially, understanding variation is key to assessing the statistical robustness of conclusions that we can draw. We ask questions such as: how reproducible are our results? and how much does the variation in the sample affect our conclusions?
14. Professor Dawson had advised that she was unable to undertake the normal organic chemistry analysis because the physical size of the samples were not large enough. However, the inorganic measurements required a much smaller sample size which allows us to take these small sub-samples (replicates) and subject each of those to analysis. The more replicates we have the more well-characterised the variability is and the more accurately and confidently we can make conclusions about the variation in the original sample. Referring to the height analogy, if I measure someone's height, I might measure it three times and say, "Well, I can measure their height to plus or minus a few millimetres". If I measure it one thousand times, I might still only be able to measure it to plus or minus a few millimetres, because that might be the accuracy of the tape measure. However, I can be much more confident that it is plus or minus a few millimetres and not plus or minus a centimetre.
15. This aforementioned limitation to Professor Dawson's analysis does not change the conclusions that can be drawn from this data set. If we had enough soil in the

Signature of witness..... 

samples for Professor Dawson to complete the organic analysis, then we would have two complementary sets of data. We would have the inorganic analysis and the organic analysis, which would help us draw clearer or stronger conclusions. If the two different independent analyses of the samples produce the same conclusions that would give us additional confidence that those conclusions are robust.

16. I have been asked what the purpose is of the open-source statistical software and data visualisation libraries. These are the tools of the trade that expert statisticians and data scientists use to analyse data and generate insights. The tools that I use are open source and freely available. Accessibility, transparency, and robustness are very important in statistical science. I chose to use the tools I used here to maximise the accessibility of the science. For example, using free and openly available tools means there is no cost barrier to reproducing my results. It also maximises transparency as people are able to see how I use those tools, understand how those tools work and what the actual algorithms underneath them are. The statistical software that was used is called 'R'. It is extremely robust and dependable, and it has been very well tested across of range of scientific and statistical disciplines for more than 20 years. It is a bedrock piece of software for modern science in a range of scientific disciplines.

Statistical Report – Section 2.2.1 Data Visualisation

17. Professor Dawson provided me with the data set I worked with, which was populated on an Excel Workbook that contained two spreadsheets. The first spreadsheet is a table of data with fifty-four samples which appear as rows in the spreadsheet, and the elemental composition percentages for all of the eleven elements summing to 100%. There were eleven columns of data in the spreadsheet plus a column of sample identifier information and a column of laboratory identifier information. The second spreadsheet is a table of sample metadata that outlined the locations and the context of each of the samples.

18. I have been asked what is the purpose of visualising the raw data? When analysing a data set, I consider it good practice to visualise the raw data set as it enables you to familiarise yourself with the data and develop an understanding of what the key characteristics of the raw data are. It helps you to spot any problems that might need to be investigated. For example, it assisted here in identifying a couple of samples that looked very different from the other samples. For these samples, the experiment may have failed or might not have worked as well as expected. In simple terms, the purpose of visualising the raw data is a mixture of a quality control check and familiarisation with the data. It helps you to think about what is in the data and how you might best go about analysing that. In this instance, I have used the Stacked Bar Plot and the Radar Chart as the two means of visualising the same data.


Signature of witness.....

Stacked Bar Plot and Radar Chart

19. The Stacked Bar Plot on page 6 and the Radar Chart on page 7 are both useful means of visualisation. They both show the elemental composition of the raw data, i.e., the raw numbers from the data set spreadsheet for each of the samples but grouped by the location and area of the sample information and coloured according to the different elements. This helps us detect patterns and shows us by eye how the compositions vary across the different replicates within an area. Additionally, it shows how areas compare to each other by highlighting where there are similarities or differences and where it looks like the experiment might not have worked very well. For each of the samples that we have, the replicates come from a given location and a given area. For example, if you look at the identifiers on the Radar Chart on page 7, you will see that we have a location which would be something like GAY016 (right boot of Sheku Bayoh) location, but then Areas 1 (toe to welt) and Area 2 (heel of sole). These are from the same location but different areas within that location. We then have different replicates within each of those different areas. Essentially, we have a three-level hierarchy of sample information; replicate -> area -> location.
20. Regarding the areas, it is interesting to note that we do not have different areas for the AM001 (left boot of PC Walker) and AM002 (right Boot of PC Walker) samples. This is attributable to the fact that there was not enough soil to take different areas effectively and as a result we only have one area within each of these two soil samples. We have two areas for GAY016 location and two for GAY017 location and for JM019 (Nicole Short's Vest) location we have three different areas.
21. It is important to note that we should be careful not to over-interpret the raw data plots at this point. This process is routinely done to gain familiarity with the data, but care needs to be taken not to over-interpret considering the complexity of the data and the fact that the data points are not independent.

What does the Stack Bar Plot show us?

22. I have been referred to page 6 of my report which depicts the Stack Bar Plot figure. The Stack Bar Plot consist of nine different panels depicting each area within each location along the x-axis. On the y-axis, we have the percentage from 0-100% where each bar is the same height, 0-100. On the x-axis we also have the individual replicate identifiers for each of the samples, and each bar is broken down into the eleven chemical composition measurements, represented by mini bars stacked on top of each other. The total height of all the bars ends up being the same (100%) but the height of the coloured bars within each segment is different. This shows the patterns of variation across the replicates. Again, caution must be applied to ensure that there is no over-interpretation as these are not independent data points. For example, if you look between the replicates and you see a decrease in the dark purple bar (Si, silicon), one of the other element

Signature of witness.....

weights must increase. Accordingly, our intuition could be a little misleading sometimes here.

What does the Radar Chart show us?

- 23. I have been referred to page 7 of my report which depicts the Radar Chart. The Radar Chart shows the data in a slightly different way. We have each of the nine mini-plots and one is depicted for each area within each location. Each mini-plot is composed of eleven different axes, all going from 0 to >60. In this instance, instead of having two axes (x and y), we now have eleven different axes, and those axes are spread out around a circle. We then plot a point on each of the axes for each element percentage for each replicate and then draw a line between them to form a shape which is colour shaded. This allows us to compare those shapes by eye which allows us to immediately identify replicates that look different. For example, the red replicate for GAY016 Area 1 looks like a different shape than the other replicates. This method makes it easy to pick out places where replicates look different or where there are similarities as our eyes are great at picking out shapes.

- 24. Both methods of visualising the data highlight where there are similarities, where there are discrepancies and whether there is anything strange that requires further investigation. For example, looking back at the Stack Bar Plot, you will see the visual discrepancy between the two samples that are GAY016 Area 1 replicate 2 and JM019 Area 3 replicate 2. The compositions are strange for those soils and are quite different from the other soils. If we do not think those samples are representative and we think that there is a problem with them, then including that additional variation by leaving them in the analysis may seriously limit the conclusions we can draw. Depending on the data set, it is an interesting statistical challenge to try to identify what looks like a real outlier and should be excluded and what is just natural variation. Initially, I did not rule out these two replicates, but later during the analysis when it became clear they have unusual characteristics, I made the decision to exclude these two as outliers.

- 25. The conclusion to exclude those two replicates was reinforced by a conversation with Professor Dawson. According to Professor Dawson, aluminium which is common to most soils which makes it a good reference. According to Professor Dawson, it is present at reasonable levels in most soils and tends to not vary strongly between soil types. Where we see a replicate with a substantial increase in the aluminium level, that indicates that something has gone wrong with the experiment rather than it being natural sample variation.

Signature of witness..... 

Statistical Report – Section 2.2.2 Data Imputation

26. I have been asked to explain the process of imputing data. On pages 7 and 8 of my report I discuss the process of data imputation. Data imputation is a process of replacing missing values in the data set with appropriate values so that the data set can then be used for further analysis. Most statistical tools that we use require data to be representative and present in all the samples. Consequently, if we see data sets with zeros or with missing data, then it is very important to try to understand what those zeros mean and what the missing data mean. Addressing the reasons for data being missing is a complex field within statistics, and there are several different approaches that can be taken.
27. As statisticians, it is imperative that we ensure that the approach or strategy that we take to deal with missing data does not introduce biases into the data set and it should be appropriate for the type of data that we are using. This requires an understanding of why the missing data is in the data set to begin with. As an example, a data point could be missing randomly. For instance, a temperature measurement sensor is set up and left to take measurements but, occasionally, some internal failure (or even something as exotic as a cosmic ray interaction) might cause it to skip or miss a measurement, effectively at random. Alternatively, the missing data might not be missing randomly; for example, the battery might run out in the temperature sensor and not be noticed until the following day. The sensor hasn't taken any measurements for this period of time because the batteries have run out, so this data isn't missing at random anymore.
28. Another type of missing data is what we call 'censored data'. This is when a measurement is taken but the measurement falls below the detection threshold of the instrument. Using our temperature sensor example, supposing our temperature sensor can only measure down to zero degrees, anything below zero degrees is just recorded as zero degrees. Accordingly, if a zero is recorded in the data, we are unable to determine whether it is actually a true value of zero, less than zero, or it could be missing data that it is just recorded as a zero. As a result, trying to understand what a zero means in the data can be a challenge.
29. In this instance, we have a technique that produces a set of measurements that form a spectrum. This spectrum is comprised of peaks corresponding to element abundances, and where we take the measurements from, and a background signal level. Any peak that is smaller than the background signal level, or even comparable to the background signal level, is difficult to measure. In these cases, the instrument then records the measurement of the peak at these points as zero. So, a zero measurement might be a true zero, but it might also be a measurement that is not zero but is less than the background.
30. I have been referred to page 9 of my report which depicts the chart where we are imputing missing data because of recorded zeros. This figure shows the patterns of zero in the data for each of the elements and how often that zero occurs. A

Signature of witness.....

white square depicts non-zero data, whereas a blue square depicts zeros. On the x-axis, we have each of the columns showing each of the eleven elements. The y-axis illustrates the pattern. Looking at the top pattern (pattern 2), you can see there are data in everything except the last three elements which are all zeros. In the next pattern (pattern 3) there is data in everything except the last four elements which are all zeros, and so on. The bar chart to the right-hand side of the grid shows you how often we see each pattern in the data set. For example, you can see that Pattern 2 occurs 46% of the time. This means that 46% of our data has non-zero measurements for the first eight elements, and zeros for the last three elements. The bar chart let you see what the frequency of each of the patterns is. The bar chart at the top then shows you how often zeros occur in the data for each of the elements, for example we can see that the manganese and copper data are 98% zeros. This means that we have very few non-zero measurements of any copper or any manganese and, in fact, the same for phosphorous.

31. This chart summarises the patterns of zeros and the frequency of zeros in the plot. This helps me to understand whether I need to impute data or whether I should just remove some data. In this case, I do not want to remove data unnecessarily because we have only a small amount of data to begin with. Censored relative compositional data that contains zeros requires imputation because we want to take logarithms of the ratios we will make, but we are unable to take a logarithm of a zero. Consequently, if we want to construct log ratios, we need to address the zeros first by dealing with them in a way that does not bias our data set and is appropriate for this particular kind of data. The R package called 'Z Compositions', does precisely this. It provides robust and reliable data imputation for left-censored - i.e., data that is censored on the low end - relative composition data. I used this package to make this plot and to impute reasonable values for the different elements. A simple way of thinking about this might be to consider taking the height measurements of ten children in a class, except I only measured nine of them; I missed one. What might be a reasonable estimate at what that tenth child's height is? One way of imputing would be to say, "Well, I don't know, so I'll just put in the average of the other nine children". So, I might take an average over the other nine children and say the tenth child has that height. However, this might not be reasonable, depending on whether I missed measuring that child at random or not. Perhaps they were missed because my tape measure did not have markings below one metre and that child was only ninety centimetres tall, and so an accurate measurement could not be taken. There are different strategies for what values to use to fill in in this case and the maths behind it is complicated. The package that I have used does robust and reliable inference on what values could fall below the detection threshold and then inserts them in a way that does not bias the data set.

Signature of witness.....

Statistical Report – Section 2.2.3 Log-Ratio Transformation

32. I have been referred to graph on page 11 and I have been asked to explain what this graph depicts. This chart illustrates a plot of the log ratios for each replicate within each location area. If you look at the top left corner, you can see that we have location area AM001 Area 3. The first panel depicts Aluminium (Al), and it shows the log ratio, using aluminium as the reference, for each of the six replicates. For the first column, it will always be exactly zero because here we are taking a ratio of the aluminium value to itself, which is always one, and the logarithm of one is zero. So, these should all be flat at zero, which they all are. The remaining columns show the replicate log ratio values for each individual element. For example, panel number two at the top is the same location area AM001 Area 3, however this is now showing the calcium (Ca) to aluminium ratio and we can see that it varies slightly between the replicates. The first row of the chart depicts one location area, the second row is the next location area, and third row is the next location area and so on.
33. What I am looking for in these mini figures is to see how stable the log ratio is between the replicates to determine if the log ratio transformation has worked well. If the replicates are consistent with each other, then they will be stable with a similar value for all the replicates. For example, looking at the silicon ratios (Si), in the second plot from the end/right, looking down the column you can see that they all appear stable. We have a slight anomaly with location area GAY016 Area 1 where one of the replicates (replicate 2) looks a bit out, and which we have identified as a potential outlier replicates from our stacked bar plots of the raw data. The values in the rest of that column are stable. Reviewing this chart provides a visual check to see if there are any problems with the log ratios and how consistent the log ratios are.
34. Wide variability on this chart would prompt me to double check my code to ensure that I hadn't made any mistake as this could be indicative of error in addition to potentially being caused by problems with the underlying data set or data for a particular element. Here replicate two of the GAY016 Area 1 samples really stands out and this is because the aluminium fraction has changed. Overall, the chart provides a consistency check.
35. I have been referred to the plot on page 12 and I have been asked to explain what this plot depicts. This plot shows the average of the replicate log ratios within each location and area. Each panel is a location area. The y-axis depicts the average log ratio value, and the x-axis shows the ten different element ratios. For example, in AM001 Area 2, the first point of the first plot shows the calcium to aluminium log ratio, the second point shows the copper to aluminium log ratio, then potassium to aluminium and so on. Together, these are log ratio signature profiles for each of the location areas. The plot allows you to review by eye and identify patterns of similarities and differences between location areas. This is a useful simplification plot as it captures how consistent the log ratios are for the

Signature of witness.....



different elements and it allows you to look at the different composition signatures. For example, looking at AM002 Area 1 and JM019 Area 2, they have a different shape to the left-hand side of their signature of this plot. This indicates that they have elemental differences in their log ratios. Whereas JM019 Area 2 shares more similarity with AM001 Area 3. I will add a caveat, which is that these are averages which are being compared by eye and as such it is not a formal quantitative comparison at this point.

Statistical Report – Section 2.3 Sample Comparison


Hierarchical Clustering

36. I have been asked to explain the concept of Hierarchical Clustering. Clustering is a broad term for a set of statistical methods that assesses the similarities between data sets, and then groups the data based on these similarities. The important characteristics of hierarchical clustering as compared to other clustering method, is that it does not make any assumptions about how many groups there are in the data or how many groups there should be. Instead, it plots the data in a hierarchy of similarity, and this is represented in a format that resembles a tree. A helpful analogy might be a family tree. For example, when people are close together on the same level in a family tree they're more closely related to each other. In this instance, we don't have family members appearing at each level, so here it is more like looking at a family tree snapshot of how a generation of people are related right now, as opposed to over time and over many generations.
37. The three plots on pages 14, 15 and 17 show the relationships between the data at all levels. In this hierarchy, the relatedness of any two samples in a plot is defined by the height of the path that you have to go along to get from one sample to the other. To identify how similar two samples are, you follow the line and see how high up the plots you need to go to get from one sample to another sample.
38. I have been referred to plots on page 14, 15 and 17, and I have been asked to explain what these plots depict. Looking at the plot on page 14, from the left-hand side, to get from AM001 Area 3 replicate 5 to GAY016 Area 3 replicate 3, you have to walk all the way up to the top of the tree and then all the way back down again to travel between these two replicates. These two replicates are well separated and not similar to each other. Conversely, looking at AM001 Area 3 replicate 5, you do not have to go far to get to GAY016 Area 1 replicate 1. These two replicates are most similar to each other.
39. There is a caveat to interpreting these plots. This plot represents a 3D structure, but presents it on a 2D plane. An example would be to think of how it would look if we made a child's hanging mobile from this tree. Then we can see that just because names of replicates that are next to each other on the x-axis of the

Signature of witness..... 

paper plot, that may not mean that they are similar to each other. Looking at the plot on page 14, AM002 Area 1 replicate 6 and GAY016 Area 2 replicate 1 are next door to each other along the x-axis. However, in this instance you will need to travel a long way up and down the tree to get between them, and if we imagine the child's mobile, in 3D these would swing apart from each other.

40. There are a lot of technical options statisticians can choose when calculating clustering, such as what way to measure the distance between things. There are lots of ways you might measure the distance between groups to produce this kind of clustering, such as using the distance between the averages of each group, or between the closest or furthest apart samples from each group. This is called the linkage. I have checked a few different ways to see if the clustering options change the clustering results substantially and I have shown an example of this on page 15 by using a different linkage for this plot, than for the plot on page 14. The plot on page 14 uses the distance between the averages of groups when clustering, the plot on page 15 uses the distance between the furthest apart samples instead. The clustering is similar in both plots and is robust to this change. As an example, if we look at the JM019 Area 3 replicates 6, 3 and 4. We see that these cluster together in both plots regardless of the choice of linkage.
41. These plots highlight the complexities of the samples because the replicates do not cluster together neatly into groups that clearly represent a single location area. This suggests that these are either uncertain or heterogeneous samples. For example, if we focus on the reference soils, which are from all the location areas that are not JM019, we can see that although sometimes they group together quite nicely, they are also often scattered across the plot. The AM002 location area replicates tend to cluster closely together as opposed to the AM001 replicates, which are on the opposite ends of the cluster tree. This shows that it is difficult to clearly distinguish between the soils based solely on this data as the replicates from the different locations and areas show similarities in their properties.
42. The conclusion that I came to from the clustering plots are that there are four broad groups. First, we have a group of replicates formed of the GAY017 Area 1 and JM019 Area 3 soil samples. We also have a group that is JM019 Area 2 mixed with a few other replicates from other location areas. The other JM019 replicates are spread throughout the clusters which shows that these questioned soil sample replicates have similarities to quite a lot of the reference sample replicates and that there is a lot of variability across the JM019 soils.
43. There are two rough groupings in the left of the plot on page 14. The left most one is predominantly replicates from AM002 Area 1. This contains all the AM002 samples but also one of the AM001 and one of the GAY016 Area 1 samples. Then we have one large group that contains everything else, mainly GAY017 Area 1, GAY017 Area 2, JM019 Area 1, a few JM019 Area 3 replicates, and

Signature of witness..... 

some of the GAY016 Area 2 replicates. So those are the four broad clustering groups.

- 44. The plot on page 17 is a much simpler clustering tree plot. In this plot, I averaged all the replicates within a location area together first and then calculated the clustering using those average values. This is much more straightforward to interpret, but we should be cautious here because this plot doesn't include any of the information about the variability of the replicates within each of those location areas, so we must apply caution not to over-rely on conclusions drawn from this simpler plot.
- 45. From the average clustering plot we can see that JM019 Area 3 is, on average, most similar to the GAY016 Area 2 and also forms a larger cluster with GAY017 Area 1 and 2. I consider this as one group. AM002 Area 1 is an outlier but is loosely related to this group. Then we have another group which splits into two sections: one is the AM001 Area 3 and GAY016 Area 1 samples. The other group is the questioned JM019 Area 1 and 2 samples. Again, it is important to note that caution should be applied to not over-interpret the average plot as this does not include any of the variation between replicates.

Statistical Report – Section 2.3.2 Principal Component Analysis- PCA

- 46. I have been asked to explain Principal Component Analysis (PCA) and what it captures. Principal Component Analysis is one of a suite of approaches that statisticians use to simplify data sets for analysis. This approach simplifies the data by reducing the number of columns that are in the data table down to a smaller number. In this instance, we have a table with ten columns, where each column is the log ratio of an element's abundance relative to aluminium, for each of our sample replicates, which are the rows. Principal Component Analysis is a way of reducing these ten columns down to fewer columns, in this case I use two. Each of the principal components, these new variables Principal Component 1 (PC1) and Principal Component 2 (PC2) – are a combination of the original ten values, with a different contribution from each of the elemental ratios in PC1 than there is in PC2.
- 47. When you go through the process of reducing the number of columns, you inevitably lose some of the information. However, it is important to note that the PCA algorithm calculates the weightings that make up the principal components in such a way that PC1 captures as much of the information from the data as it can. Then the next principal component, PC2, captures as much of the information as it can that has not already been captured by PC1. In this instance PC1 and PC2 combined capture just over 70% of the information in the original log ratio data.
- 48. These two new columns make it easier to identify the similarities between samples in the data whilst retaining as much of the information as possible. In



Signature of witness.....

summary, PCA is used to simplify the data set but maintain as much of the information as possible, so that we can draw some clear conclusions.

- 49. I have been referred to plot on page 19, and I have been asked to explain what this plot illustrates. The x-axis depicts PC1, and the y-axis depicts PC2. The data points show the PC1 and PC2 values for each individual replicate. On this plot, points that are close together represent samples which are similar to each other. Conversely, points that are far apart relate to samples which are not similar to each other. PCA provides both a quantitative and visual measure of similarity.
- 50. There are three tiers of information in the data set that is represented on the plot, these are: the location, area and replicate. Each of the replicate data points are colour coded according to the location and area they originates from. This is represented in the legend shown on the right side of the plot, where you see each of the location areas separated by a colour. Each of the data points are then labelled with a number indicating the replicate id within that location area. The plot allows us to identify and capture the patterns of clustering in the data; where replicates are similar to each other and where location areas are similar to each other.
- 51. For this plot, I used a paired colour scheme because the locations all have two areas (with the exception of the JM019 locations). I chose this to allow the viewer to clearly identify similar locations. For example, if you look at both of the AM locations AM001 and AM002, they are different shades of blue. Both of the GAY016 locations are coloured in different shades of green. This colour system helps people identify all three tiers of information that we have for these samples.
- 52. The most complicated part of the plot are the coloured ellipses which are drawn behind the data points. There are indicative areas of the plot that are occupied by, or that represent, each location area. The shape and size of the ellipse is defined by the variability of the replicate data within each of those location areas. In the report, I say these are the one standard deviation ellipses for the multivariate normal distributions of the data. In simple terms, I have defined these ellipses so that if you imagine we had one thousand data points from each location area, I expect about 68% of the data points to lie inside of that ellipse. This means that not all of the data points are expected to lie within the ellipse. The 68% size of the ellipse is essentially arbitrary, and they are there only as a guide to illustrate the relative scale of the variability and as an indication of area of the principal component plot that is occupied by replicates from that location area.
- 53. An important quantity is the middle of the ellipse as this defines the middle of that location area. For example, GAY016 Area 1 and AM001 Area 3 ellipses are at slightly different angles, and each has a large spread of their replicates. However, the centres of those two ellipses are pretty close to each other and they both overlap. This reinforces our previous clustering plot where AM001 and GAY016 Area 1 showed some significant similarities, because here they occupy

Signature of witness..... 

a similar space in the plot. It is important to note that there is a lot of variation between these replicates, however, and as such we would be unable to draw any strong conclusions about the similarity of these locations even though their centres are close. This highlights the heterogeneities in the sample. In some cases, we may have samples that are mixed. In simple terms, imagine taking two different soils and mixing them together as one sample. This would naturally increase the amount of variation that you see when you then take replicates from that sample.


54. There are a few replicates that overlap each other which has made it difficult to distinguish the numbers on these plots. This is evident in GAY017 Area 2 replicates 4, 9 and 2. This overlap indicates that these replicates are very similar to each other. Soils are a heterogeneous medium, so they have quite a lot of variation, but where we see things clustering tightly together, this gives us confidence that that area is really well defined. If those are then different or similar to, or in the same area of the plot or a different area of the plot than replicates from other samples, we can make more robust conclusions about them.

Conclusions:

55. I have been referred to page 19 of my report where I state:

“The JM019 Area 1 replicates are not tightly clustered, and do not consistently cluster with a distinct set of other samples, suggesting this sample is highly heterogeneous and may not be from a single origin”

I have been asked to clarify what this means. We can see that for JM019 Area 1, we have six different replicate measurements that are all spread out along both PC1 and PC2. The PCA analysis produces an ellipse that is long and looks similar to a cigar shape with replicates 5 and 6 being in the bottom left of the plot. Replicates 1, 4 and 3 are near the upper middle of the plot and replicate 2 being somewhere in the middle. This spread probably represents a soil that is very heterogeneous (which means that it has a lot of variation in its characteristics) or it could be that there are actually several soils mixed together. For example, imagine getting dirt on your shirt and then getting another bit of dirt on your shirt in the same place and then taking samples which would have those two bits of dirt mixed together. There is no straightforward way of determining whether this variability is a result of heterogeneity or whether it is a mixed sample. However, replicates 5 and 6 line up with GAY016 Area 1 and AM001 Area 2 at their Principal Component 1 values, whereas replicates 1, 4 and 3 line up with some of the other samples and this is certainly consistent with being a mixed sample.

Signature of witness.....

56. I have been referred to page 20 of my report where I state:

“The JM019 Area 2 replicates cluster tightly, and are well separated from the other samples suggesting a separate origin for this soil”

I have been asked to clarify what this means. The ellipse for JM019 Area 2 is considerably smaller than that of JM019 Area 1. Most of the replicates, with the exception of replicate 3 cluster tightly together. They are also well separated from the other samples and they do not show a strong overlap with any of the other samples. They are closest with JM019 Area 1 replicate 2 or with AM002 Area 1 replicate 3. If we are looking at the central point of the ellipse, the ‘centroid,’ of AM002 Area 1 and the centroid of JM019 Area 2, they are well separated and do not show a strong relationship. The centroid of the ellipse for JM019 Area 2 is close to the centroid of the ellipse for JM019 Area 1, however, the replicates are spread apart which indicates that they come from the same location area but that they’re not necessarily from the same soil and that the relationship with JM019 Area 2 is weak

57. I have been referred to page 19 of my report where I state:

“The JM019 Area 3 replicates 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”.

I have been asked to clarify what this means. JM019 Area 3 is best represented on the zoomed plot on page 19. The JM019 Area 3 replicates all group together closely. The ellipse is small as the replicates cluster tightly, overlap and occupy the same area of the plot as the replicates of GAY017 Area 1 and 2 and GAY016 Area 1. JM019 Area 3 also shares the same area as JM019 Area 1 replicates 1, 4 and 3 which represents the similarities between all of these different replicates.

58. I have been asked to clarify the difference between homogeneous and heterogeneous, in this context. A homogenous soil would be a soil sample where replicate sub-samples taken from the soil all show similar characteristics. This means that the soil is well-mixed and uniform which would lead to tight replicate clustering on this PCA plot. Conversely, a heterogeneous soil sample is one that is very different in different places, which means that replicate sub-samples will show a wide range of characteristics and hence will be more spread out on the PCA plot. In simple terms, the more tightly clustered replicates are together, the more homogeneous they are. The more spread out they are, the more heterogeneous they are. My conclusion for the JM019 Area 3 samples from the Nicole Short’s vest is that this is consistent with sharing a common original with the soils GAY017 soils and GAY016 Area 2 soils from Mr Bayoh’s boots.

Signature of witness..... 

59. I have been referred to page 19 of my report where I state:

“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.”

I have been asked to clarify what this means. GAY016 Area 2 and GAY017 Area 1 overlap strongly with separation between replicates within each. GAY016 Area 2 replicates 1, 5 and 6 are well separated from replicates 2, 3 and 4 along Principal Component 1 creating two sub-groups of replicates. There is a similar pattern of separation within GAY017 Area 1 where replicates 1, 2 and 3 are separated from 4, 5 and 6, and where 4 and 6 overlap. Both GAY017 Area 1 and GAY016 Area 2 shows the same sub-group structure. This reinforces that these two soils are not only similar but that they have the same kind of clustering sub-structure to their replicates.

Statistical Report – Section 2.3.2.1 Principal Component Loading

60. I have been referred to the table on page 20 of my report and I have been asked to explain Principal Component Loadings and how this has affected my PCA results. I have referred to this process when discussing the contribution, (also referred to as weighting or loading) of the log ratios that makes up each of the principal components. Each of the principal components is a combination of all of the ten elemental log ratios, but how much each of those log ratios contributes to each principal component is different, which is what is shown in this table. For example, looking at the first row of the table, we can see the contribution of the log ratio of sodium (Na) to aluminium (Al), to each principal component. The sodium log ratio contributes more to PC1 than it does to PC2.

61. This table is used to show the patterns of information captured by each principal component. The principal components loadings have a minimum value of -1 and a maximum value of 1. For example, if we look at magnesium (Mg) the PC2 is 0.58 which is quite a big positive contribution. It is also a lot more than its 0.07 contribution to PC1. We can see from the table that the sodium (Na), sulphur (S), copper (Cu) and manganese (Mn) ratios are important for PC1. Whereas the magnesium, (Mg) sodium (Na) and calcium (Ca) ratios are more important for PC2. Ideally, a principal component would not be completely dominated by one column, and instead would have a mix of contributors with different weightings. This would show that all of the different elements are important for characterising the soil but that they operate in different patterns across the principal components.




Signature of witness.....

Statistical Report – Section 2.4 Quantifying the probability of Soil Sample Relationships

62. I have been asked to explain the concept of a threshold-based null hypothesis test. A hypothesis is a possible explanation we have about a scientific problem or observation that we want to test. In this instance, our hypothesis is that two samples that we want to compare are different from each other. For every hypothesis that we generate there is a corresponding 'null hypothesis'. In this instance, the null hypothesis is that the two samples are similar to each other and are consistent with coming from the same source or parent sample.
63. To be more precise, our hypothesis is that the two sets of replicate samples are consistent with coming from different origins. Whereas our null hypothesis is the opposite; our null hypothesis is that the two sets of replicate samples come from the same origin or are consistent with coming from the same origin. The 'p-value' is statistical measure that is commonly used for null hypothesis testing. The p-value is the probability of obtaining data as extreme, or more extreme, as the observed data, by chance, if your null hypothesis is true. For example, if your hypothesis is that two samples are different and you perform a statistical test to quantify how different they are, you can use a p-value to assess whether the test results allow you to reject your null hypothesis. Importantly, it does not allow you to confirm that your hypothesis is actually true, just whether or not you can reject the null hypothesis.
64. The way that a threshold hypothesis test works is that, prior to undertaking any of the calculations, you must set a threshold value for your p-values. This threshold set the level at which you are prepared to reject the null hypothesis. In this instance, we use a threshold p-value of 0.05, which represents a 1-in-20 chance of us rejecting the null hypothesis when it is actually correct. If the p-value is lower than 0.05 then we say that we reject the null hypothesis.
65. It is important to note that there are some subtleties around p-value calculations, particularly when you are performing lots of tests. This is typically addressed by what is called 'multiple testing correction' to prevent over-interpretation of the results. For example, suppose we have a hypothesis that eating jelly beans causes acne, and I do twenty tests and in one of those tests the p-value is below our threshold of 0.05. If I only report the positive result that fell below the p-value threshold then we might conclude that there is a real effect, but actually because I have done twenty tests with a p-value threshold of 0.05, we might expect to get one positive result even when the null hypothesis is true. In summary, whenever you conduct a lot of tests, you should use a multiple testing correction of your p-values to prevent mis-reporting.

Conclusions

66. The plots on pages 23, 25 and 26 represent the p-value results from the statistical test that I used, which is a version of the Kolmogorov-Smirnov test (KS-

Signature of witness..... 

test). This test has the advantage of looking directly at the data without assuming anything about the distributions of the data involved. Other statistical tests may make assumptions about whether the data is normal or not. The KS-test is a conservative test and is less sensitive compared to some other tests due to the lack of assumptions about the data. The dotted line on the graphs is the multiple testing corrected p-value threshold we are using. This method reinforces the picture that we have collated from the other clustering results in a quantitative way.

67. The conclusions that are drawn from the threshold-based null hypothesis testing are consistent with the Principal Component Analysis and Hierarchical Clustering results. The null hypothesis testing is performed using the principal components rather than on the log ratios themselves, whereas the hierarchical clustering and the Principal Component Analysis use the log ratios. The fact that these different analyses paint the same picture reinforces the conclusions.


68. I have been referred to page 23 of my report where I state:

“We can reject the null hypothesis for AM001 Area 3 and both GAY017 areas and thus conclude that:

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

I have been asked to clarify what this means. The plot on page 23 shows the p-value and the threshold for JM019 Area 1 compared to each of the other reference locations. On the x-axis we have each of the reference locations and on the y-axis, we have the multiple testing corrected p-value. The dotted red line indicates the multiple testing corrected p-value threshold. This plot shows that we can reject the null hypothesis for AM001 Area 3 and both GAY017 soils. This means that these samples are consistent with having a different origin i.e., not sharing a common origin with JM019 Area 1. Conversely, we cannot reject the null hypothesis for AM002 Area 1 and the two GAY016 soils. This means that for those three areas the null hypothesis may be correct and so they are consistent with sharing a common origin with JM019 Area 1.

Taking a conservative approach, I interpret above and below the threshold strictly without looking at the scale of the numbers involved. This conclusion supports the picture that we saw from the Hierarchical Clustering and the Principal Component Analysis. A less conservative approach that takes into account the values of the p-values, not just whether they are above or below the threshold, would conclude that, additionally, JM019 Area 1 is more consistent with sharing a common origin with GAY016 1 and GAY016 2 than it is for AM002 Area 1, even though we're unable to formally reject the null hypothesis for AM002 Area 1 at the threshold level.

Signature of witness.....  ..

69. I have been referred to page 25 of my report where I state:

“We can reject the null hypothesis for all the soils and conclude that the JM019 Area 2 soils is not consistent with sharing a common origin with any of the soils examined here.”

I have been asked to clarify what this means. The plot on page 25 shows the p-value threshold for JM019 Area 2. On the y-axis we have the multiple testing corrected p-value and on the x-axis, we have the reference soils that we are comparing with JM019 Area 2. We can see that all of the data points fall below the threshold, although the error bar for AM002 Area 1 spans the threshold. Again, taking a conservative approach, I reject the null hypothesis for all the reference location areas, including AM002 Area 1, and conclude that JM019 Area 2 is not consistent with sharing a common origin with any of the reference soils

Again, it is important to note that null hypothesis testing only tells you whether you are able to reject the null hypothesis or not, given the data. It does not comment on whether the hypothesis is true or not. For example, if we look at the same graph plot on page 25, it does not tell us whether JM019 Area 2 is from AM001, it only comments on whether we can reject the null hypothesis.

70. I have been referred to page 26 of my report where I state:

“We can reject the null hypothesis for AM001 Area 3, AM002 Area 1, GAY016 Area 1 and GAY017 Area 2 and thus conclude that:

- *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.*
- *JM019 Area 3 sample is consistent with sharing a common origin with GAY016 Area 2 and GAY017 Area 2 soils.*

I have been asked to clarify what this means. The plot on page 26 shows the p-value threshold for JM019 Area 3. On the y-axis we have the multiple testing corrected p-value and on the x-axis, we have the reference soils that we are comparing with JM019 Area 3. We can reject the null hypothesis for all the soils except for GAY016 Area 2 and GAY017 Area 1. It is interesting to note that there is significant variability within the data and that there are some instances where the error bar crosses the threshold line. Consequently, we only just reject the null hypothesis for these cases.



Signature of witness.....

Statistical Report – Section 2.5 Conclusions and Caveat

71. As far as already not explained, I have been asked to explain the conclusions set out on page 26 of my report:

- *“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.”*

These conclusions are drawn from integrating the results from each of the different part of my analysis, so they are not dependent wholly on one part of the analysis over any of the others.

72. The first conclusion refers to JM019 Area 1, which I have concluded is from a heterogeneous sample that has similarities to several of the known reference samples in this analysis. Based on the replicates, the soil sample is most similar to GAY016 Area 1 and GAY016 Area 2. These were the replicates that were tightly embedded with those areas on the PCA plot. It is notable that both of these samples come from a similar location which is the right boot of Mr Bayoh, and the null hypothesis testing indicates this sample is consistent with sharing a common origin with these soils. It is important to note that a few of the replicates from JM019 Area 1 are similar to some of the other replicates due to the heterogeneity of the soil. This soil may be a mix of soils that come from several locations and several of those locations may be the reference locations.

73. The second conclusion refers to JM019 Area 2. I have concluded that this is a relatively homogeneous sample which means that the replicates are fairly similar to each other. They are not as variable as the JM019 Area 1 replicates but there is still a significant amount of variability, and they are distinct from the other samples in this examination. The null hypothesis testing rejects all the reference soil location areas as sharing a common origin, they occupy a different space on the PCA plot, and they cluster separately from most of the other replicates. The origin is unknown which means that this could have come from somewhere else. We do not have any information to draw any further conclusions.

Signature of witness.....


74. The third conclusion relates to JM019 Area 3. The JM019 Area 3 replicates are the strongest piece of evidence. I have concluded that they are a homogeneous sample, in that the replicates cluster tightly together. They have similarities to several of the known reference samples in the examination. They are most similar to and consistent with sharing a common origin with GAY016 Area 2 and GAY017 Area 1, which are soils from the left and right boots of Mr Bayoh. It is notable that they are also similar to some replicates of GAY017 Area 2 which is an area from a similar location.


Comment on Findings:

75. I have been asked whether my findings support or contradict the findings of Professor Dawson's report. My findings support Professor Dawson's findings which were reached independently. There are some minor differences between our findings because there are some things that I cannot say based on the data that Professor Dawson can from her analysis. For example, Professor Dawson can rule out AM002 as an origin from JM019 Area 1 based on the size and shape of the particles in the soil sample that she sees under the microscope. I do not have that information. I only have the elemental composition information. Other than a few minor differences, the major conclusions are supported.

Comment on Probability and Likelihood:

76. I have been asked to explain why I was unable to comment on the statistical likelihood or probability of the soil samples sharing a common origin. The normal way we would go about calculating likelihoods is to use an alternative proposition approach, for example, you would have two alternative propositions and you would be comparing one to the other. This would then allow us to make statements like the probability of this soil coming from location A is 10 times greater than the probability of coming from location B. We are unable to define the absolute probability of a soil coming from a given location without some alternative propositions. For alternative propositions, we might try to use something like the National Soils Database as a set of reference alternative propositions and use these to calculate likelihoods. However, soil is a very variable medium and varies substantially over small distances, so this is not ideal.

77. In this instance we do not have appropriate reference data as there were no reference soil samples taken in the area, and additionally Lorna and I were specifically instructed by the inquiry not to take an alternative proposition approach for the soils samples that were collected and analysed here. So, for example, I was asked not to assess the likelihood or probability of the questioned samples coming from Mr Bayoh's boots compared to PC Craig Walker boots.

Signature of witness..... 

Miscellaneous

78. The purpose of this statement to is to provide a summary of the key elements of my statistical analysis and the conclusion of my report.

79. I believe the facts stated in this witness statement are true. I understand that this statement may form part of the evidence before the Inquiry and be published on the Inquiry's website.

December 2, 2022 | 3:59 PM GMT

Date..... Signature of witness.....



Signature of witness.....