A Data-Driven Investigation Into Similarity Measures For Global Chemical Products

Matthew Watson

MSc by Research

University of York

Mathematics

April 2022

ABSTRACT. Croda is one of the largest chemical companies in the United Kingdom, producing and distributing products across the globe. It is the aim of this research to provide Croda a means for determining the similarity and therefore interchangeability of their products between manufacturing sites. To do this, numerous analytical approaches including the Bhattacharyya distance, Mahalanobis distance, hierarchical clustering, distribution modelling and separation are investigated. Novel approaches to outlier detection and exploratory analysis are also examined. These analyses are applied to three data sets, each corresponding to a chemical product - Tween20, BrijCS20 and Glycerox HE. These data sets consist of the mass charge ratios and their abundance obtained via MALDI-TOF mass spectrometry.

Of the analyses conducted, hierarchical clustering as well as distribution fitting yielded the most promise, although both methods were susceptible to outliers. The Gaussian model, for example, fits the data for the products quite accurately but is less accurate for higher masses. In conclusion, it is found that finding the desired similarity measure is extremely challenging.

Contents

List Of Figures

List Of Tables

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

1. Introduction

Headquartered in the United Kingdom, Croda International plc is a major chemicals company manufacturing a range of speciality chemical products sold across the globe [**1**]. Originally founded in 1925 by George Crowe and Henry Dawe, Croda has grown rapidly from a struggling lanolin producer to one of the most profitable chemical companies in the UK [**2**]. As such, Croda now has manufacturing sites scattered around the world, which were previously small, independent chemical companies with their own recipes for products. Generally, Croda have not standardised recipes across sites after acquisition and have kept the production process unchanged at the acquired sites. As a result, some sites will have slightly different recipes (ingredients used, ratio of ingredients and catalysts used are different) and clearly this can lead to differences in the final products between sites. Another possible cause for this discrepancy between products from different sites is likely due to Croda's ethical environmental policy of sourcing product ingredients local to each site [**3**]. For example, animal product used in some of Croda's products is always sourced local to each factory and not from one central location [**3**]. Hence, the aim of this project is to develop a similarity measure which can successfully determine the "sameness" (similarity) of products between the Croda manufacturing sites. This will allow us to determine which sites are interchangeable for certain products, and help us to understand why differences occur. With this aim in mind, a number of similarity measures and data analysis techniques including the Bhattacharyya distance, the Mahalanobis distance, curve fitting (Gaussian, Weibull and Fréchet), and hierarchical clustering are investigated.

The data from which this investigation is conducted on is that obtained by matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI-TOF-MS) of three chemical products from across various Croda manufacturing sites. Previous work including [**4**] and [**5**] suggest that it is possible to find practical similarity measures for chemical data of this type. Therefore, we hypothesise that it is plausible and desirable to develop such a measure for this data; however, this optimistic view must be cautioned with the caveat that data, especially mass-spectrometry data, is often notoriously challenging [**6**] [**7**].

2. Experimental Methods

In this project, the MALDI-TOF-MS data of three Croda chemical products manufactured globally are analysed; these products are Tween20, BrijCS20, and Glycerox HE. Tween20 is the brand name for polysorbate20, and it consists mostly of fatty acids (precisely: stearic, palmitic, myristic, and lauric acids) [**8**] [**9**] . This product has many commercial uses; in particular, as an oil in water emulsifier, solubiliser for essential oils and perfumes, and wetting agents [**9**]. Tween20 is a complex product with approximately 20 different compounds involved in its composition, which will allow for method testing. BrijCS20, on the other hand, is an alkyl polyglycol ether, and is made by reacting a fatty alcohol mixture of stearyl and cetyl alcohol (and some other minor components) with 20 moles of ethylene oxide [**10**]. Primarily found in hair and skin care products, this simpler product, with just two main constituents, acts as dispersing agents, emulsifiers and surfactants / detergents [**11**], and will be used to facilitate method development. Finally, Glycerox HE is a mixture of glycerin, coconut oil and approximately 21 moles ethylene oxide [**12**]. It is used in an array of skin and hair care products, acting as oil in water emulsifiers, superfatting agents, solubilisers and as dispersing agents. This product is simpler than both Tween20 and BrijCS20, and therefore provides additional data for method development and testing.

As aforementioned the method of data acquisition is the soft ionisation technique MALDI-TOF-MS. Mass spectrometry is regarded as one of the best methods for accurately determining the molar mass of molecules [**13**]. Accurately measuring molar mass is the best-known approach for identifying molecules (and their amount) in a substance. All molecules and substances have mass (measured in grams); however, these masses are extremely small. Therefore, the SI base unit, the mole, which is a measure of the amount of a substance is needed for comparing particles of a substance and its mass. When the number of moles is known, the concept of molar mass can be applied to calculate the number of grams of the substance. Hence, Molar mass, measured in grams per mole, is the total mass of all atoms in a mole of a molecule. Mathematically then, the molar mass M is simply given by:

$$M = \frac{m}{n}$$

where m is the mass of the sample substance in grams, and n is the number of moles of the sample substance being analysed.

During mass spectrometry, the samples are vaporised (transformed into a gaseous state) and then bombarded with electrons in a process known as ionisation. The mass/charge (m/z) ratio can then be calculated for all ions. Mass/charge ratio is an extremely important measure in chemistry which allows one to find the relative abundance of an element or molecule in a compound. This will be important in our analysis since it will be used to determine the chemical composition of samples of products. One major challenge which arises during traditional mass spectrometry techniques is the fragmentation of large molecules during vaporisation; MALDI-TOF-MS, however, largely overcomes this challenge. Another useful property of MALDI-TOF-MS from an analytical viewpoint is that the charge is always one, so we do not need to be concerned with charge.

In MALDI-TOF mass spectrometry, the samples to be analysed are embedded in a solid matrix, [1] vaporised and then ionised. These ions are then accelerated across an electric field for a distance d_E, and then drift across a region for a distance d_D before reaching a detector. Ions with the smallest m/z ratio will reach the detector before those with larger m/z ratios.

[1] In chemistry, this refers to the surface upon which samples are placed in mass spectrometry.

An ion of mass m undergoing MALDI-TOF mass spectrometry will have a charge of ze, where z is the charge number of the ion, and e is the elementary charge (electric charge carried by a single proton). This ion accelerates across an electric field of strength E for a distance d_E, and then drifts a distance d_D, until it arrives at the detector. Therefore, the kinetic energy E_K of the ion is given by:

$$\begin{aligned} E_K &= \frac{1}{2}mv^2 \\ &= zeEd_E \end{aligned}$$

where v is the speed, in metres per second (ms^{-1}), of the ion.

Since the time of flight t for an ion to travel from the solid matrix to the detector and the distance d_D are sufficiently small, the acceleration can be ignored, so that $v = d_D t^{-1}$. Substituting this into eqn. 1, we obtain,

$$\tfrac{1}{2}m\tfrac{d_D^2}{t^2} = zeEd_E$$

Which can be rearranged to give the mass/charge ratio:

$$\tfrac{m}{z} = 2eEd_E\tfrac{t^2}{d_D^2}$$

In this experiment, each sample is analysed three times (these are technical replicates) so that the consistency (and variation) of the mass spectrometry analysis can be assessed. This means, for example, that if a product has 50 samples then it has 150 replicates (3 for each sample).

Example (Example MALDI-TOF Mass Spectrum). Figure 1 shows the mass spectrum (relative abundance against m/z values) obtained by MALDI-TOF-MS of a sample of BrijCS20 from the Atlas Point manufacturing site. Clearly, two completely overlapping distributions, corresponding to the two major components within the product, are visible in Figure 1. These distributions are related peaks, which are 44 Daltons apart in their m/z values, of the same compound. The reason for the two different distributions is that the EO units (number of ethylene oxide molecules) added to each is different. That is, the spectral peaks correspond to the same compound but are polymers of different lengths. The peaks show the relative abundance of their corresponding m/z values; that is, the height of a given peak represents the relative abundance of the polymer associated with it and is proportional to the abundance of said peak. It appears that these distributions may be approximately Gaussian; this intriguing observation is investigated in detail in section 5.6.

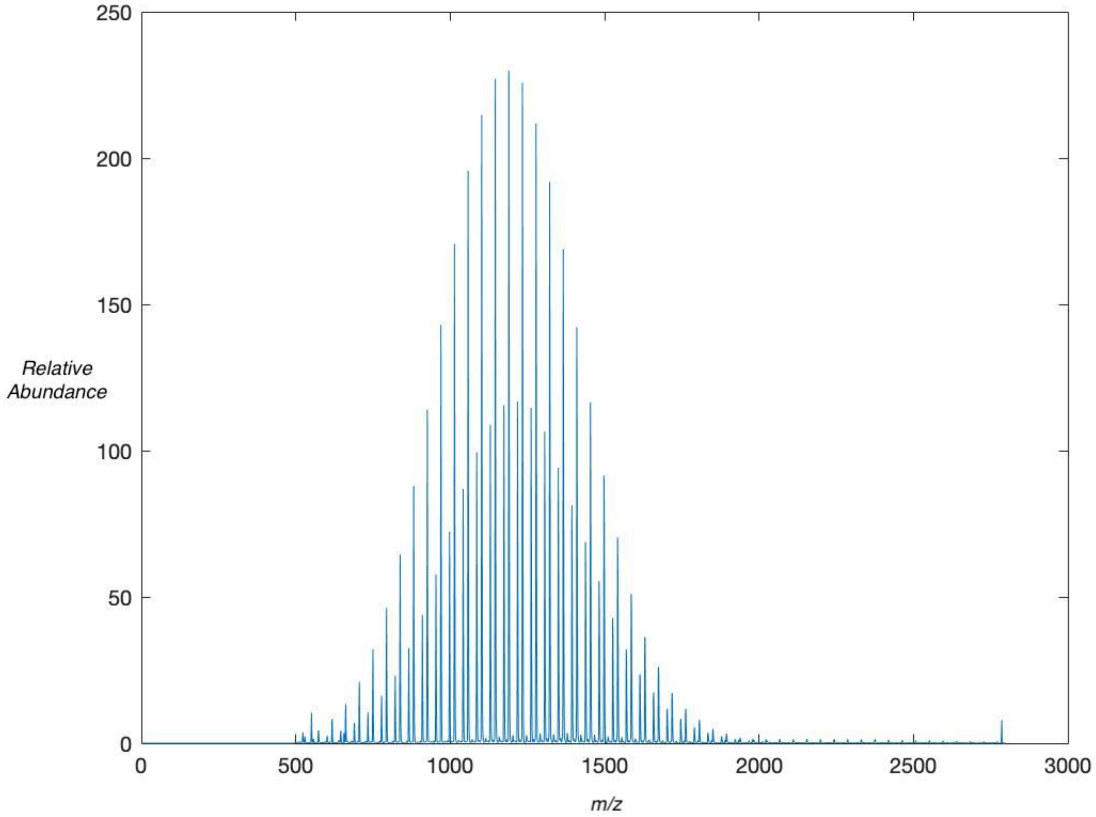


Figure 1. **MALDI-TOF Mass Spectrum of a sample of BrijCS20 obtained from the Atlas Point Manufacturing Site**. Distributions of related peaks of the same compound from BrijCS20 are shown as an example of a mass-spectrum.

3. Data

There are three data sets in this analysis, one for each product.

For Tween20, the raw data consists of 2100 variables (the masses obtained from MALDI-TOF-MS) and 48 observations from across six Croda sites. Two of these sites are located in the USA: Atlas Point and Mill Hall; two are in Europe: Rawcliffe Bridge (UK) and Chocques (France); and two are in Asia: Singapore and Thane (India).

With BrijCS20, the raw data consists of 2800 variables and 50 observations. In this data set the BrijCS20 product is produced across five Croda sites: Rawcliffe Bridge, Singapore, Thane, Atlas Point and Mevisa (Spain).

Raw data for Glycerox HE consists of 1200 variables and 50 observations. This data set is constructed from observations from four Croda sites: Atlas point, Mill Hall, Rawcliffe Bridge, and Singapore.

A useful summary of the data is provided in Table 1.

Table 1. Table 1 shows the number of observations from each site for the Tween20, BrijCS20 and Glycerox HE products. In total there are 48 observations of Tween20, 50 of BrijCS20, and 39 of Glycerox HE acquired from various Croda sites. Each site producing BrijCS20 has provided 10 observations. For Tween20 two sites, Atlas Point and Rawcliffe Bridge provided less observations (4 each) than the rest (10 observations each). Glycerox HE has 10 observations for each site except for Atlas Point which has 9.

Site	Samples of Tween20	Samples of BrijCS20	Samples of Glycerox
Atlas Point	4	10	9
Chocques	10	–	–
Mevisa	–	10	–
Mill Hall	10	–	10
Rawcliffe	4	10	10
Singapore	10	10	10
Thane	10	10	–

4. Statistical Methods

4.1. Important Definitions.

Definition 1. A *distance metric* (simply a *metric* and sometimes a *distance function*) is a function d which outputs a distance between each pair of observations of a set X. For a distance to be classed as a metric, a collection of well-defined axioms must be satisfied. That is, a metric d on X is a function $d : X \times X \to [0, \infty)$, if for all $x, y, z \in X$ axioms A1-to-A3 hold:

A1 $d(x, y) \geq 0$ and $d(x, y) = 0 \iff x = y$ non-negativity
A2 $d(x, y) = d(y, x)$ symmetry
A3 $d(x, y) \leq d(x, z) + d(z, y)$ triangle inequality

Definition 2. A *metric space* is a set which has a metric. [**14**]

Euclidean and squared Euclidean distance are perhaps the most common and straightforward similarity measures used in cluster analysis, as well as many other branches of mathematics. The Euclidean distance is a metric, whereas the squared Euclidean distance is not since it does not satisfy the triangle inequality; it is, however, still a useful distance measure. The Euclidean distance between two vectors $\mathbf{x} = (x_1, x_2, x_3, ..., x_n)$ and $\mathbf{y} = (y_1, y_2, y_3, ..., y_n)$ is denoted $d(\mathbf{x}, \mathbf{y})$ and is given by:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

Derivation of 2-Dimensional Euclidean Distance:

Consider two points **A** and **B** with coordinates (x_1, y_1) and (x_2, y_2) respectively.
Let d be the distance between them, so that $d = \text{dist}(\mathbf{A}, \mathbf{B})$.
Join the points **A** and **B** by a line segment.
Construct a right angled triangle with the line segment joining A to B as the hypotenuse. The points A and B join a point C as shown in Figure 2 to form the right angled triangle ABC.

Applying Pythagoras' theorem to triangle ABC gives:

$$\begin{aligned} AB^2 &= AC^2 + BC^2 \\ d^2 &= (x_2 - x_1)^2 + (y_2 - y_1)^2 \\ \implies d &= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \end{aligned}$$

Remark. Not taking the final step gives the squared Euclidean distance.

Proof. *Euclidean Distance is a Metric*

To prove that the Euclidean distance is a metric, we must show that it satisfies all axioms A1, A2 and A3 from definition 1.

Proof of A1 (non-negativity):

By definition

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

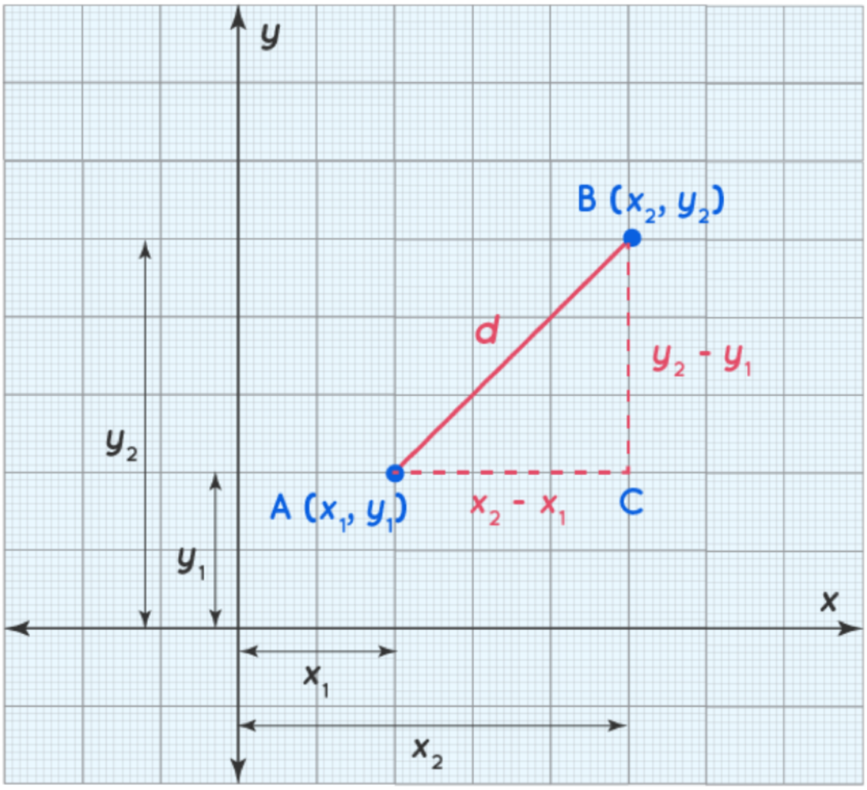


FIGURE 2. **Diagram to Illustrate the Derivation of the 2-Dimensional Euclidean Distance.** The right angled triangle ABC with hypotenuse d is shown. In order to illustrate the derivation of the 2-dimensional Euclidean Distance, the length of the line segments (excluding the hypotenuse) of the triangle are shown in terms of their coordinates of the points A, B and C.

where $x_k, y_k \in \mathbb{R}$.

Firstly, since $(x_k - y_k)^2 \geq 0$, we have $\sum_{k=1}^{n} (x_k - y_k)^2 \geq 0$ and therefore $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \geq 0$.

Secondly, when $\mathbf{x} = \mathbf{y}$ we have:

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= d(\mathbf{x}, \mathbf{x}) \\ &= \sqrt{\sum_{k=1}^{n} (x_k - x_k)^2} \\ &= \sqrt{0^2} \\ &= 0 \end{aligned}$$

Similarly for $d(\mathbf{x},\mathbf{y}) = 0$; if $\sum_{k=1}^{n}(x_k - y_k)^2 = 0$, then $(x_k - y_k)^2 = 0$ which implies that $x_k = y_k \ \forall \ k = 1, 2, ..., n$. Hence, $d(\mathbf{x},\mathbf{y}) = 0 \Longleftrightarrow \mathbf{x} = \mathbf{y}$ and A1 holds.

Proof of A2:

$$\begin{aligned} d(\mathbf{x},\mathbf{y}) &= \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} \quad \text{by definition.} \\ &= \sqrt{\sum_{k=1}^{n}(y_k - x_k)^2} \\ &= d(\mathbf{y},\mathbf{x}) \quad \text{by definition} \end{aligned}$$

Hence, A2 holds.

Proof of A3 is beyond the scope of this work. □

For the comparison of distances, it is often more practical to use the squared Euclidean distance because it removes the square root step and thus reduces the cost in time complexity of calculation. Squaring the distance can also give more weight to the effect of larger distances between observations.

DEFINITION 3. The *squared Euclidean distance* for two vectors $\mathbf{x} = (x_1, x_2, x_3, ..., x_n)$ and $\mathbf{y} = (y_1, y_2, y_3, ..., y_n)$ is defined by the formula:

$$d^2(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{n}(x_k - y_k)^2$$

PROOF. *Squared Euclidean distance is not a metric by counter-example.*
Let $\mathbf{y} = 2\mathbf{x}$ and $\mathbf{z} = 3\mathbf{x}$ for some real-valued non-zero vector $\mathbf{x}$.
Substituting into the triangle inequality gives:

$$\begin{aligned} d(\mathbf{x},\mathbf{y})^2 &= \sum_{i=1}^{n}(x_i - 3x_i)^2 \\ &= 4\sum_{i=1}^{n}x_i^2 \end{aligned}$$

and

$$\begin{aligned} d(\mathbf{x},\mathbf{y})^2 + d(\mathbf{y},\mathbf{z})^2 &= \sum_{i=1}^{n}x_i^2 + \sum_{i=1}^{n}x_i^2 \\ &= 2\sum_{i=1}^{n}x_i^2. \end{aligned}$$

∴ we have a counter-example that demonstrates that the triangle inequality does not hold, since $4\sum_{i=1}^{n}x_i^2 > 2\sum_{i=1}^{n}x_i^2$. Hence, it is shown by counter-example that Squared Euclidean distance is not a metric. □

The Manhattan distance is another commonly-used distance metric. It is different to the Euclidean distance and is classed as a taxicab metric. This metric works by calculating the absolute difference between coordinate pairs in a data set, and is defined in definition 4.

Definition 4. The *Manhattan distance* between two real-valued vectors $\mathbf{x} = (x_1, x_2, x_3, ..., x_n)$ and $\mathbf{y} = (y_1, y_2, y_3, ..., y_n)$ is defined as:

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= ||\mathbf{x} - \mathbf{y}|| \\ &= \sum_{i=1}^{n} |x_i - y_i| \end{aligned}$$

4.2. Principal Component Analysis. Principal Component Analysis[2] (PCA) is a dimensionality reduction technique, which is often used when handling large sets of multivariate data. This is because large data sets can be (and usually are) difficult to explore and analyse; since visualising such data can become near-impossible when the number of variables involved is too large. PCA aims to overcome this challenge by constructing new variables, called principal components (PCs), which are uncorrelated, linear combinations of the original variables. These principal components provide a far smaller, and thus practical, set of underlying or characteristic variables whilst retaining enough information to describe the data accurately. Ideally then, it is desirable for the few first principal components to account for the majority of the variance in the analysis. However, it is important to note that even if this is the case, it does not necessarily result in variables that can be interpreted.

Remark. PCA is classed as an unsupervised technique since it does not assume anything about the groupings in the data.

Derivation of Principal Components via Eigenvector Decomposition:

This derivation is based on the derivation found in [**16**], which in turn is based on Hotelling's approach [**17**].

Consider the random vector $\mathbf{x} = x_1, ..., x_n$ as the set of original variables and let the random vector $\mathbf{y} = y_1, ..., y_n$ be linear combinations of the original variables such that

$$y_i = \sum_{j=1}^{m} a_{ij} x_j, i = 1, ..., n$$
$$\text{or } \mathbf{y} = \mathbf{A}^T \mathbf{x}$$

where $\mathbf{A}$ is the matrix of coefficients.

We want to find the $\mathbf{A}$ generating new variables, y_j, with stationary values of their variance; that is, with constant variance over time.

The first new variable is given by

$$y_1 = \sum_{j=1}^{m} a_{1j} x_j$$

Choosing $\mathbf{a}_1 = (a_{11}, a_{12}, ..., a_{1n})^T$ maximises the variance of y_j under the constraint that $\mathbf{a}_1^T \mathbf{a_1} = |\mathbf{a_1}| = 1$. Hence,

$$\begin{aligned} var(y_1) &= E[y_1^2] - E[y_1]E[y_1] \\ &= E[\mathbf{a}_1^T \mathbf{x}\mathbf{x}^T \mathbf{a}_1] - E[\mathbf{a}_1^T \mathbf{x}]E[\mathbf{x}^T \mathbf{a}_1] \\ &= \mathbf{a}_1^T (E[\mathbf{x}\mathbf{x}^T] - E[\mathbf{x}]E[\mathbf{x}^T])\mathbf{a}_1 \\ &= \mathbf{a}_1^T \sum \mathbf{a}_1 \end{aligned}$$

[2] Karl Pearson originally introduced the concept of principal component analysis in his 1901 paper: *On Lines and Planes of Closest Fit to Systems of Points in Space*, which can be found in reference [**15**].

where $\sum$ denotes the covariance matrix of $\mathbf{x}$.

To find the stationary value of $\mathbf{a}_1^T$ such that $|\mathbf{a_1}| = 1$ we solve an equivalent problem, which is to find the stationary value of

$$\mathbf{a}_1^T \sum \mathbf{a}_1 - \mathcal{L}\mathbf{a}_1^T \mathbf{a}_1$$

where $\mathcal{L}$ is a Lagrange multiplier. Then by equating to zero and differentiating with respect to the components of $\mathbf{a}_1$ we have $\sum \mathbf{a}_1 - \mathcal{L}\mathbf{a}_1 = 0$. We want to find a solution, other than the null vector, for $\mathbf{a}_1$. Therefore, we want to find the eigenvector of $\sum$ with the eigenvalue $\mathcal{L}$. The n eigenvalues, $\lambda_1, ..., \lambda_n$, associated with $\sum$ can be ordered such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0$. The variance of $\mathbf{y}_i$ is $\mathcal{L}$, so we want $\mathcal{L} = \lambda$ to maximise variance as desired. Therefore, the first PC will be y_1 and will have the greatest variance of all the PCs. We can then obtain the 2nd PC $\mathbf{y}_2 = \mathbf{a}_2^T \mathbf{x}$ as follows: select a_{2i} for $i = 1, ..., n$ such that $var(\mathbf{y}_2)$ is maximised under the constraints $|\mathbf{a_2}| = 1$ and $\mathbf{y}_2$ and $\mathbf{y}_1$ are uncorrelated. Therefore, $E[\mathbf{y}_2\mathbf{y}_1] - E[\mathbf{y}_2]E[\mathbf{y}_1] = 0$ or, equivalently, $\mathbf{a}_2^T \sum \mathbf{a}_1 = 0 \implies \mathbf{a}_2^T \mathbf{a}_1 = 0$ ($\mathbf{a}_2$ and $\mathbf{a}_1$ are orthogonal) since $\mathbf{a}_1$ is an eigenvector of $\sum$. To maximise we use the Lagrange multipliers $\mathcal{L}_1$ and $\mathcal{L}_2$ so that $\mathbf{a}_2^T \sum \mathbf{a}_2 - \mathcal{L}_1 \mathbf{a}_2^T \mathbf{a}_2 - \mathcal{L}_2 \mathbf{a}_2^T \mathbf{a}_1$. Differentiating with respect to $\mathbf{a}_2$ and setting to zero we obtain: $2\sum \mathbf{a}_2 - 2\mathcal{L}_1 \mathbf{a}_2 - \mathcal{L}_2 \mathbf{a}_1 = 0$. Then multiply by $\mathbf{a}_1^T$ to get: $2\mathbf{a}_1 \sum \mathbf{a}_2 - \mathcal{L}_2 = 0 \implies \sum \mathbf{a}_2 = \mathcal{L}_2 \mathbf{a}_2$. Hence, $\mathbf{a}_2$ is an eigenvector of $\sum$ too, and is orthogonal to $\mathbf{a}_2$. It is the eigenvector corresponding to the second largest eigenvalue. This argument repeats and extends to the k-th case.

Thus, we can determine the PCs via eigenvector decomposition. We have that $\mathbf{y} = \mathbf{A}^T \mathbf{x}$, where the columns of the matrix $\mathbf{A}$ correspond to the eigenvectors of $\sum$.

Next, to produce an accurate dimensionally reduced representation of a given data set, we observe that $\sum_{i=1}^{n} var(y_i) = \sum_{i=1}^{n} \lambda_i$ and then that the first k PCs account for $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$. Since the new components have to account for a minimum percentage of the total variance, say p, then k is chosen such that $\sum_{i=1}^{k} \lambda_i \geq p \sum_{i=1}^{n} \lambda_i \geq \sum_{i=1}^{k-1} \lambda_i$ and transform the data such that $\mathbf{y_k} = \mathbf{A_k^T x}$.

4.3. Normalising Data. To ensure that the data is found on a consistent range of values for each observation, it is common practice to normalise the data [**18**]. To do this, the sum of variables for each observation in the data set are set to the same value.

4.4. Squared Mahalanobis Distance Measure. One similarity measure investigated in this analysis is the Squared Mahalanobis distance metric [**19**]. It is a popular data-driven similarity measure and is commonly used for cluster analysis to evaluate the distance between two points (or group means) [**19**]. As such, it is a good candidate for a similarity measure for this data.

The Squared Mahalanobis distance between two points $\mathbf{x} = (x_i, x_j)$ and $\mathbf{y} = (y_i, y_j)$ with covariance matrix $\mathbf{S}$ in 2-dimensional space is given by:

$$D_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{y})$$

The sample covariance matrix $\mathbf{S}$ can be used to determine the correlation between data points. This is extremely useful because any important correlation between variables of data points which may otherwise have seemed quite distant from each other can be identified with this measure.

Furthermore, taking into account the Within Groups variance, the Squared Mahalanobis distance between two group means $\mathbf{C} = (c_i, c_j)$ and $\mathbf{E} = (e_i, e_j)$ is given by:

$$D^2_M(\mathbf{C}, \mathbf{E}) = (\mathbf{C} - \mathbf{E})^T \mathbf{S}^{-1} (\mathbf{C} - \mathbf{E})$$

where the sample covariance matrix $\mathbf{S}$ is calculated as the mean of the covariance matrices of each group. That is, for groups i and j

$$\mathbf{S} = \frac{(n_i-1)\mathbf{S}_i + (n_j-1)\mathbf{S}_j}{n_i+n_j-2}$$

Where n_i and n_j are the number of observations from groups i and j, respectively.

4.5. Bhattacharyya Distance Measure. Another distance metric that could potentially act as a successful similarity measure for this data is the Bhattacharyya distance metric. This metric is used to measure the relative closeness (similarity) of two probability distributions [**20**]. It is a common distance measure applied to multivariate data (examples include the many works by Ron Wehrens such as [**21**] amongst others) and it is an extension of the multivariate Mahalanobis distance and therefore is a measure worth investigating. This measure assumes the data is normally distributed, and in the multivariate case is given by the equation:

$$D_B = \frac{1}{8}(\mu_2 - \mu_1)^T \sum{}^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{det(\sum)}{\sqrt{det(\sum_1) det(\sum_2)}}$$

where, for $i = 1, 2$, μ_i and $\sum_i$ are respectively the mean and covariances for distribution i, and $\sum = \frac{\sum_1 + \sum_2}{2}$.

4.6. Within Groups Variance, Between Groups Variance and Separation. Before detailing Within Groups Variance, Between-Groups Variance, and Separation it is important to first define some key concepts as follows.

Definition 5. The *Sample mean* is the best-known measure for identifying the centre of a data set and is computed with the formula:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where n is the number of observations in the data set.

Example (Calculating the Sample Mean). Consider the heights, in metres, of the last five different Olympic Mens 100m gold medallists:

$$1.88, 1.95, 1.85, 1.76, 1.85.$$

$$\begin{aligned}\therefore \quad \overline{x} &= \frac{x_1 + x_2 + x_3 + x_4 + x_5}{n} \\ &= \frac{1.88 + 1.95 + 1.85 + 1.76 + 1.85}{5} \\ &= 1.858 \quad \text{metres.}\end{aligned}$$

Similarly, using the heights, in metres, of the last five unique Olympic Womens 100m gold medallists, we obtain:

$$\begin{aligned}\overline{x_2} &= \frac{1.67 + 1.52 + 1.73 + 1.6 + 1.7}{5} \\ &= 1.644 \quad \text{metres.}\end{aligned}$$

Definition 6. The *Sample Variance* provides a means of quantifying the variance across the observations in a data set. It is defined by the formula:

$$S_n^2 = \tfrac{1}{n-1} \textstyle\sum_{i=1}^{n} (x_i - \overline{x}_n)^2$$

Example (Calculating the Sample Variance). Using the same data set as that in example 2, the sample variance for the Mens 100m gold medallists is:

$$\begin{aligned}S^2 &= \frac{1}{n-1}((x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + (x_4 - \overline{x})^2 + (x_5 - \overline{x})^2) \\ &= \frac{1}{5-1}((1.88 - 1.858)^2 + (1.95 - 1.858)^2 + (1.85 - 1.858)^2 + (1.76 - 1.858)^2 + (1.85 - 1.858)^2) \\ &= 0.00467 \quad \text{metres squared.}\end{aligned}$$

Similarly, for the womens data,

$$\begin{aligned}S_2^2 &= \frac{1}{4}((1.67 - 1.644)^2 + (1.52 - 1.644)^2 + (1.73 - 1.644)^2 + (1.6 - 1.644)^2 + (1.7 - 1.644)^2) \\ &= 0.00713 \quad \text{metres squared.}\end{aligned}$$

Definition 7. The *Grand mean* is the mean of means of groups within a data set, and is defined by:

$$X_{GM} = \frac{\sum\limits_{i=1}^{k} n_i \overline{x}_i}{\sum\limits_{i=1}^{k} n}$$

for unequally sized groups and

$$X_{GM} = \tfrac{1}{k} \sum_{i=1}^{k} \overline{x}_i$$

for equally sized groups.

Where k is the number of groups within the data set, n_i is the sample size of group i and x_i is the sample mean for group i.

Example (Calculating the Grand Mean). Using the same data set, the grand mean for the Male-Female 100m runner groups is:

$$\begin{aligned}X_{GM} &= \frac{\overline{x} + \overline{x_2}}{k} \\ &= \frac{1.858 + 1.644}{2} \\ &= 1.751\end{aligned}$$

As the name suggests, Within Group variance is a measure of the variance within individual groups in a data set. This statistic is often used in the analysis of variance (ANOVA) and has the general formula:

$$Var_{WG} = \sum_{j=1}^{g} \sum_{i=1}^{k_j} \frac{(x_{ij} - \overline{x}_j)^2}{k_j - 1}$$

Similarly, the definition of between groups variance is self-evident; it is a measure of the variance between distinct groups. It achieves this by quantifying how the group means differ from one another. This is also commonly used in ANOVA and is calculated using the following general formula:

$$Var_{BG} = \sum_{j=1}^{g} \frac{n_j(\overline{x_j} - \overline{X}_{GM})^2}{g - 1},$$

where g is the number of groups.

Separation is a measure of similarity derived from both within and between groups variance. It is simply calculated as:

$$\text{Separation} = \frac{\text{Between-Groups variance}}{\text{Within Groups variance}}$$

4.7. Hierarchical Clustering. Hierarchical clustering analysis (HCA) is a commonly used [**22**] method for analysing clusters of multivariate data. As the name suggests, HCA works by creating a hierarchy of clusters from data. There are two main approaches to HCA, namely: Agglomerative, which is a bottom-up method, and Divisive clustering, which is a top-down method. In the agglomerative approach, each observation is initially its own cluster (from the full spectrum of product data); pairs of clusters then merge to form new clusters as they move up the hierarchy. Divisive clustering is the opposite approach, where there is initially one cluster consisting of all observations, this single cluster then splits as it moves down the hierarchy.

In this analysis, only agglomerative hierarchical clustering is implemented for the creation of dendrograms (discussed later in this section) and for the measure of distance and similarity. This is because divisive clustering is computationally far less efficient than agglomerative clustering for multivariate data [**24**]. As such, agglomerative clustering is used frequently and divisive is not for such data [**24**]. Thus, we will only use agglomerative clustering in this analysis.

The first step in agglomerative hierarchical clustering is to link single observations into a cluster, and then merge to other clusters and so on up the hierarchy. To determine the distance between an observation and the members of the other clusters, a linkage function is employed. There are a number of different linkage methods that could be used; however, four stand out as the most widely used and practiced in cluster analysis [**22**]. As such, this analysis is restricted to just these four and are detailed as follows:

1. Single Linkage

Single linkage [22] clustering takes the minimum distance between two observations, one from each cluster, as the distance between them. This linkage function is also known as the nearest neighbour method and is defined as:

$$D(X,Y) = min_{x \in X, y \in Y} d(x,y)$$

where X and Y are two clusters, and x and y are elements of X and Y respectively. One important criticism of this linkage method is that long thin clusters are often generated since observations further away in a cluster are not evaluated.

2. Complete Linkage

Complete linkage [22] is another easily understood linkage method. Here, the distance between two observations is the maximum distance between them. It is also known as the farthest neighbour linkage method and is defined as:

$$D(X,Y) = max_{x \in X, y \in Y} d(x,y)$$

Generally, this method produces tighter and more compact clusters (often approximately equally sized) than single linkage [22].

3. Average Linkage

The average linkage [22] method involves calculating the mean distance of all possible pairs of observations between two clusters. Also known as the unweighted pair group method (UPGMA), this method is defined as:

$$D(X,Y) = \frac{1}{N_X, N_Y} \sum_{x \in X} \sum_{y \in Y} d(x,y)$$

where N_X and N_Y is the size (total number of observations) of cluster X and Y respectively.

An important remark is that this is an unweighted method, and therefore all the calculated distances contribute in an equal proportion to the final average.

4. Ward's Method

First presented by Joe H. Ward in 1963, Ward's linkage method [23] is an algorithm which recursively minimises the within group variance. As such, clusters are formed according to variance. This is a distinctly different approach compared to the other three linkage methods discussed since, unlike those, Ward's method does not use a distance measure for linkage.

The distance between clusters by Ward's method is the squared Euclidean distance between points:

$$d_{ij} = d(\{X_i\}, \{X_j\}) = \|X_i - X_j\|^2$$

Ward's method often leads to clusters of roughly equal size and is highly sensitive to outliers [22].

Once HCA is complete and the numerical output is obtained, it can then be represented diagrammatically using a dendrogram. This is often more informative since it provides a clear and concise picture of how separate the data clusters are. A dendrogram consists of branches and leaf nodes. The branches in a dendrogram join clusters together; the height at which they merge represents the distance between observations or clusters; while, the leaf nodes are the individual observations themselves.

Example. **Dendrogram using the inbuilt USArrests data in R.**
The USArrests data set in R consists of the arrests per 100,000 in each of the 50 States in the year 1973. The percentage of the population registered as living in urban regions is also included. From this data the following dendrogram (Figure 3) is produced:

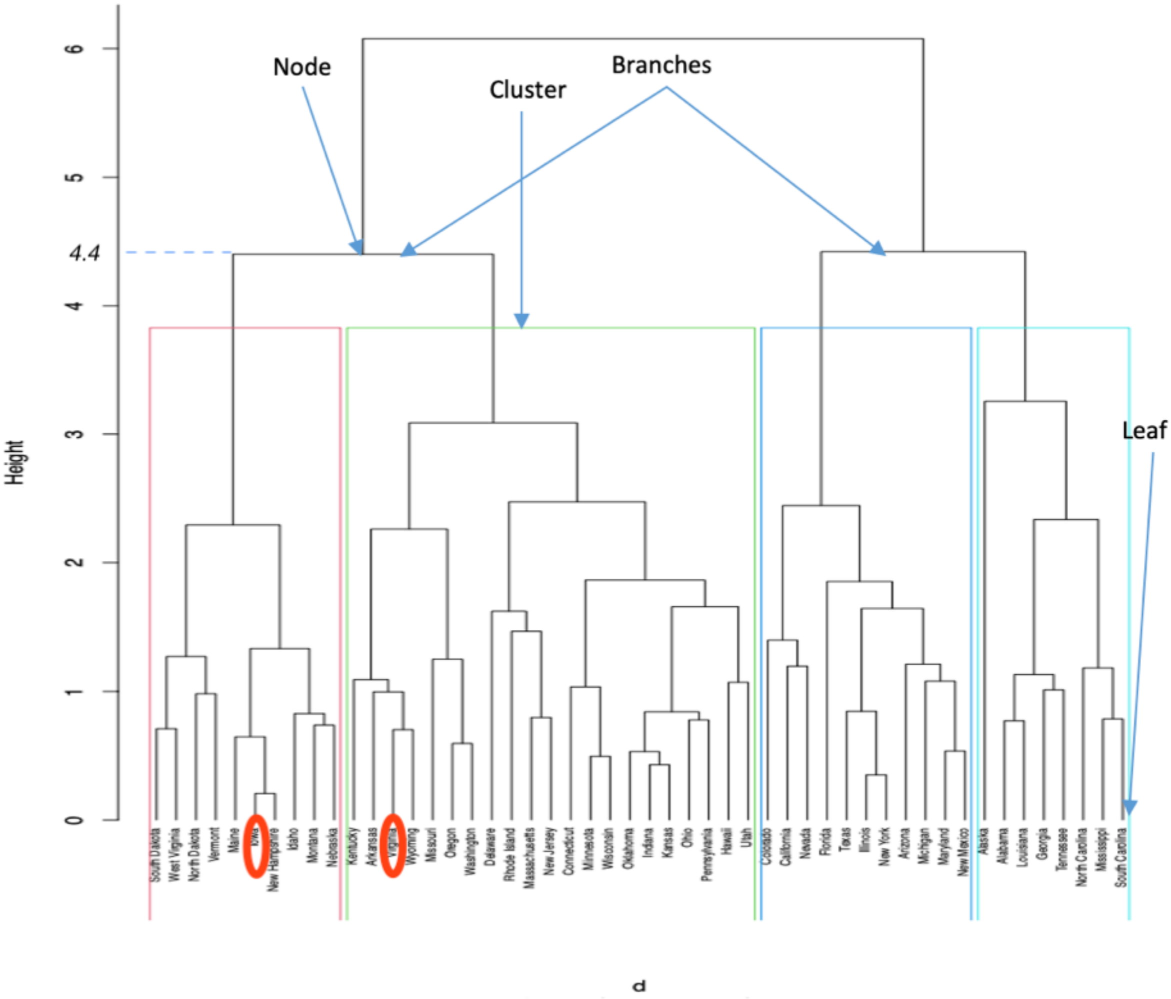


Figure 3. **Illustration of a Dendrogram using the inbuilt USArrests data set in R**. Dendrogram using Euclidean distance for USArrests data which consists of the arrests per 100,000 in each of the 50 States in the year 1973. The percentage of the population registered as living in urban regions is also included. The diagram highlights the key features of a dendrogram - nodes, clusters, branches and leaves.

The height of the dendrogram where two observations, and therefore two branches, join to form a cluster is known as the cophenetic distance (dissimilarity between observations). For example, the cophenetic distance between “Iowa” and “Virginia” in

Figure 3 (see observations circled in red) is approximately 4.4 as shown. A distance matrix and a cophenetic dissimilarity matrix are therefore created from the clustering algorithm. To measure the extent to which the dendrogram accurately represents the original data, the cophenetic correlation coefficient is used. We will combine two different approaches of calculating the cophenetic distance in order to derive a practical similarity measure. We have the asymmetric (and thus not a metric) cophenetic distance and the symmetric version which uses the average Within Groups similarity (so can generate values greater than 1). To obtain the similarity measure we simply invert the values obtained in the cophenetic matrices; that is, similarity = 1/dissimilarity = 1/(cophenetic distance). The asymmetric element of the measure tells us how similar the products are between manufacturing sites, whereas the symmetric element is used only for within sites measurements - it gives us an indication of how consistent the product is internally for each manufacturing site.

4.8. Methods for Detecting Outliers.

4.8.1. *Boxplots.* Perhaps the most common approach when attempting to identify outliers is to use a boxplot. Invented in 1970 by John Tukey [**25**], the box-and-whiskers plot as it is formally known provides a visual summary of the spread of ones' data.

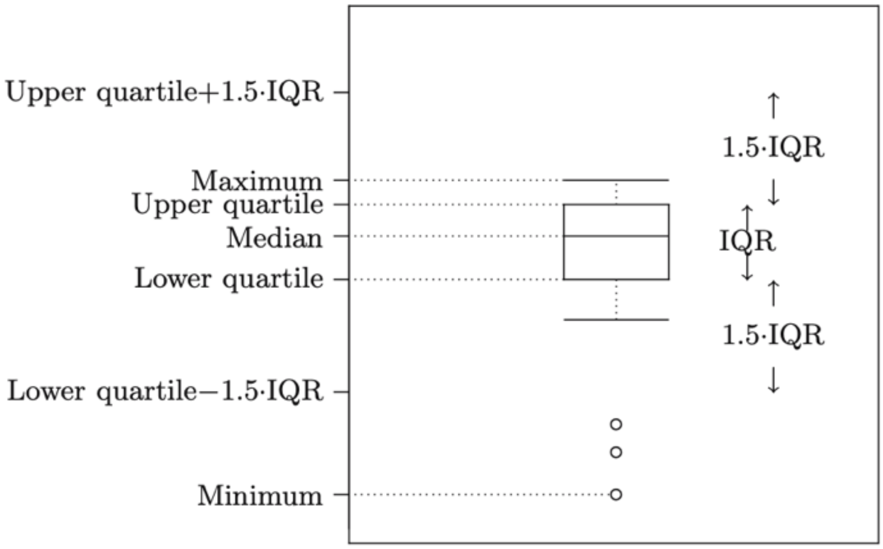


FIGURE 4. Boxplot taken directly from Figure 16.3 from page 237 of "A Modern Introduction to Probability and Statistics" by F.M. Dekking et al (Springer). [**26**]

In this work (see section 5.5.1) outliers are classed as observations found outside the fences defined by $w_1 = q_1 - 1.5(q_3 - q_1)$ and $w_3 = q_3 + 1.5(q_3 - q_1)$, or when appropriate $w_1 = q_1 - 3(q_3 - q_1)$ and $w_3 = q_3 + 3(q_3 - q_1)$, where q_1 is the lower quartile and q_3 is the upper quartile. Figures 14 and 15 in section 5.5.1 include both "inner" (green) and "outer" (red) fences calculated using these formulae. These formulae are chosen since they are commonly used for outlier identification [**26**] and can be visualised easily on a boxplot.

4.8.2. *Bagplots.* First published in 1999 by Rousseeuw, Ruts and Tukey, the Bagplot [**27**] (occasionally called starburst plot) is a bivariate extension to the univariate boxplot (also Tukey). It provides a visualisation of the spread and nature of bivariate data, and is used for identifying outliers in such data sets. In Figure 5, up to half of the data is found in the dark blue segment, known as the bag. In the construction of this bagplot, and all others, there is also a fence, so in total the bagplot has three nested polygons. Although the fence is never actually plotted it is used to construct the bagplot. Observations, highlighted as red lines, between the bag and the fence are marked by a light blue segment, known as the loop. Anything that goes beyond the loop is classed as an outlier. The asterisk in Figure 5 represents the depth median; that is, the point where the Tukey depth is highest.

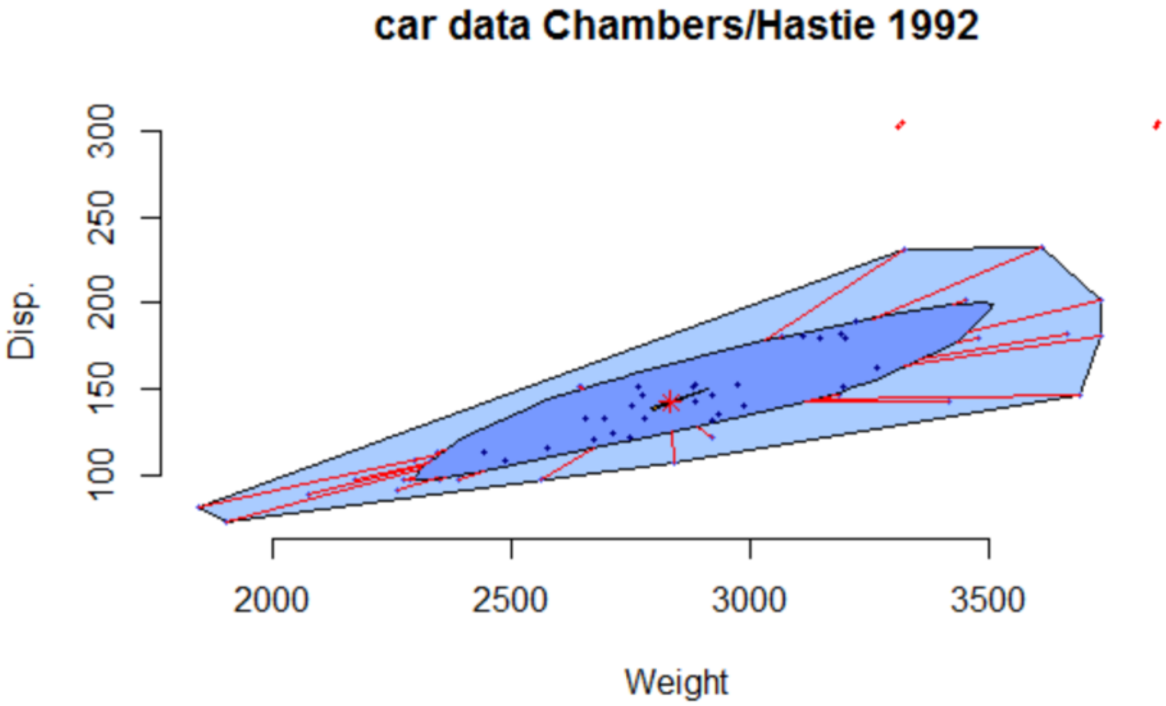


FIGURE 5. **An Example of a Bagplot.** A Simple Bagplot showing weight against displacement for car data Chamber/Hastie 1992, taken directly from [**28**].

REMARK. Code for bagplots obtained and adapted from Rousseeuw, Ruts and Tukey's original work (see[**27**]).

4.9. Distributions and Curve Fitting. A number of possible distributions were considered for modelling the three data sets, with the likeliest best-fits (and practical) distributions investigated and discussed in this report. These distributions include Gaussian (described as below), Weibull and Fréchet (these were considered and experimented with but it soon become clear that these distributions were not suitable and thus further details have been omitted).

The Gaussian[3] or normal distribution is one of, if not, the most important distributions in probability and statistics. This distribution is described by two parameters: the mean μ and the standard deviation σ. Formally, we can define the Gaussian by its probability density function f:

Definition 8. If a continuous random variable follows a *Gaussian distribution* parameterised by μ and σ, then it must have the probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for } x, \mu \in \mathbb{R} \text{ and } \sigma \in \mathbb{R}_{0+}.$$

To calculate the parameters μ and σ a variety of methods are examined. One approach is to calculate μ using the following algorithm:

For each distribution:

1. Find the three m/z values with the highest intensities. Call these I_0, I_1, I_2, where I_0 has the largest intensity and I_1, and I_2 have, respectively, the next largest, and are found either side of I_0 in the distribution plots (such as those in Figure 1).
2. Calculate $I_0 - I_1 = \text{diff1}$ and $I_1 - I_2 = \text{diff2}$.
3. If diff1/diff2 ≥ 1 then take MZ corresponding to I_0 as μ. That is, $\mu = MZ_{I_0}$.
4. Else if diff1/diff2 < 1 then take the MZ value corresponding to the midpoint between I_0 and I_1 as μ. That is, $\mu = \frac{MZ_{I_0} + MZ_{I_1}}{2}$.

Then, to estimate σ for each distribution a more complex procedure is required. To do this we first calculate the Full Width at Half Maximum (FWHM) for each of the selected distributions. The FWHM is simply the width of a peak at half its height. For mass spectrums and distributions such as ours the FWHM is a commonly used concept (see [**30**], for example) and provides a means of estimating parameters, in this case σ, since the extrema of such distributions tend to be heavily affected by noise whereas the central bulk does not.

If our data does in fact follow a Gaussian then we can use extracted features like the FWHM along with maximum height of the distribution, the ratio between two distributions' maximum heights and the centre of the distribution as variables in further analyses. This is extremely beneficial since these extracted features are clearly more practical and easy to understand than 100s of M/Z values. Using extracted features - a smaller set of variables that accurately summarise given data - is common practice and PCA is in fact an example of this.

The extracted features mentioned above are used in section 5.5 (The Hierarchical Clustering Approach) and in section 5.6 (Distribution Modelling).

Figure 6 illustrates FWHM and other important features of the Gaussian distribution.

[3] Named after the great German Mathematician and Physicist Carl Friedrich Gauss, who made one of the first applications of the distribution in 1809 [**29**]

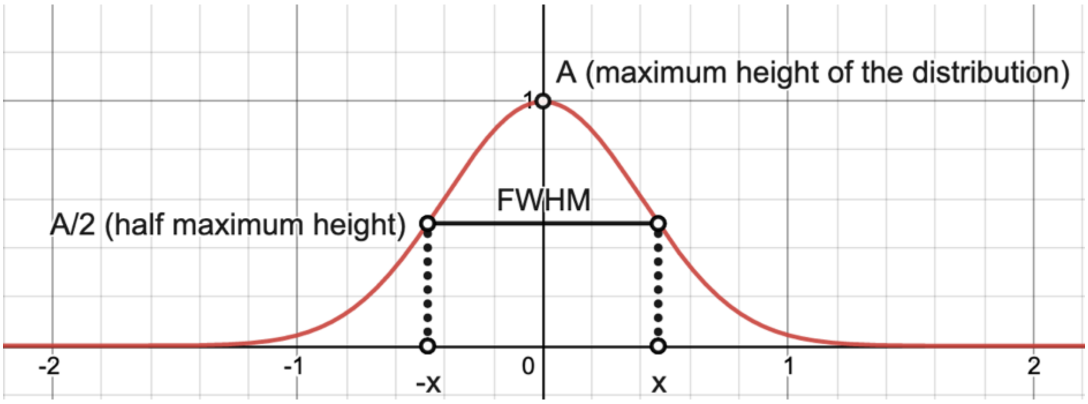


FIGURE 6. **Illustration of Full Width at Half Maximum of a Gaussian Distribution**. Figure 6 illustrates the FWHM of a Gaussian curve with height A.

FWHM is calculated as follows:

For each distribution:

(1) Find the maximum intensity in the distribution and call this H.
(2) Define the constant $h = \frac{H}{2}$ as the half height of the distribution.
(3) Find the first and last peak in the distribution with an intensity greater than h, and record their corresponding MZ values. Call these MZ values P_1 for the first peak and P_2 for the last peak.
(4) Find the intersection of the straight line joining the intensity of P_1 and the intensity of the peak 44 Daltons to the left of it, with the line $y = h$. Similarly, for P_2, find the intersection of the straight line joining its intensity and the intensity of the peak 44 Daltons to the right of it, with the line $y = h$. Call these points of intersection A and B. The corresponding MZ values for the points A and B (call these MZ_A and MZ_B) are then recorded. The diagrams in Figure 7 illustrate the concept.

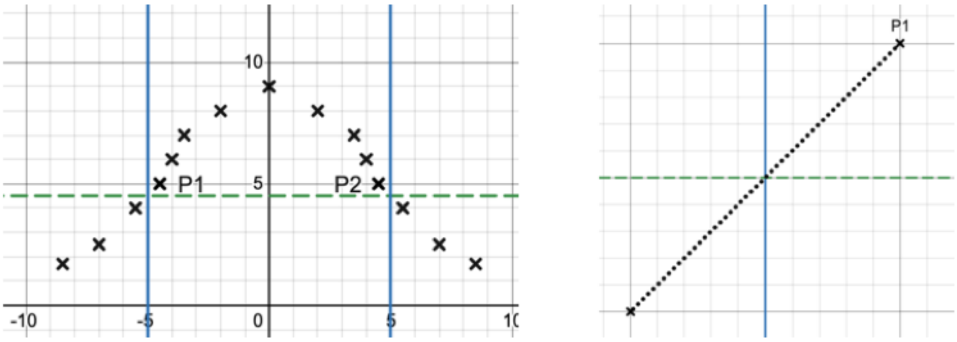


FIGURE 7. **Illustration of the Intersection of Points for FWHM Calculation**. The left diagram shows the peak points for the actual data and how the intersection between the two points is more accurate than the difference between P1 and P2, which would give a narrower FWHM than it should be. The right diagram shows how a straight line can be used to find the intersection with the line y = h, and how the FWHM can be calculated more accurately.

(5) FWHM is calculated as the difference between MZ_B and MZ_A. That is, FWHM = $MZ_B - MZ_A$.

The standard deviation σ is derived as follows:

The probability density function of a Gaussian distribution can be expressed as:

$$f(x) = Aexp\left(\tfrac{-x^2}{2\sigma^2}\right)$$

Where, A is the maximum height of the distribution.

The mean μ is set to 0.

Remark. This is legal as a Gaussian is a symmetric distribution and the nature of the distribution does not change by shifting it (and thus its mean) along the x-axis. That is, its position changes, but its nature and shape does not.

At half maximum height of the distribution, $f(x) = \frac{A}{2}$. Therefore, we have

$$\begin{aligned} \frac{A}{2} &= Aexp\left(\frac{-x^2}{2\sigma^2}\right) \\ \implies \frac{1}{2} &= exp\left(\frac{-x^2}{2\sigma^2}\right) \end{aligned}$$

By taking the natural logarithm and applying the laws of logarithms, we obtain

$$\begin{aligned} \ln\frac{1}{2} &= -\frac{x^2}{2\sigma^2} \\ \implies -\ln\frac{1}{2} &= \frac{x^2}{2\sigma^2} \\ \implies \ln 2 &= \frac{x^2}{2\sigma^2} \\ \implies x &= \pm\sqrt{2\sigma^2 \ln 2} \end{aligned}$$

Since the Gaussian curve is symmetric, the FWHM, which is simply the distance between x and $-x$ (as demonstrated in the diagram - Figure 6) is clearly given by

$$\begin{aligned} \text{FWHM} &= 2|x| \\ &= 2\sqrt{2\sigma^2 \ln 2} \\ &= 2\sigma\sqrt{2\ln 2} \end{aligned}$$

Finally, rearranging we see that

$$\sigma = \tfrac{FWHM}{2\sqrt{2\ln 2}},$$

which is used to calculate the standard deviation for each distribution.

All these parameter values are stored as a matrix from which the Gaussian curves are fitted to each distribution.

5. Results and Discussion

In this section the results of the analysis are discussed in detail. The first step in the analysis is the exploratory data analysis stage, where PCA and normalising the data are investigated using the techniques described in sections 4.2 and 4.3. This initial phase of the analysis provides a visual representation of, and insight into, the data from which we can progress from. It also allows us to determine whether or not the data should be normalised. Once the exploratory analysis is complete we investigate various similarity measures, starting with two commonly used measures: the Squared Mahalanobis distance (see 4.4 for details) and the Bhattacharyya distance (described in section 4.5). Then, by using the Within-Groups and Between-Groups variances, the Separation (described in section 4.6) is investigated as a measure. A novel approach which employs the use of hierarchical clustering and cophenetic distances is also examined (methods for hierarchical clustering are discussed in 4.7). The final approach examined in this work is to use a Gaussian fit to model the data and use extracted features (see section 4.9) from this model in a potential similarity measure. Of course, the accuracy of the fit is crucial to this approach. In summary, this section provides discussion on the validity of all these measures and analyses used to determine such.

5.1. Exploratory Analysis. For the raw Tween 20 data, PC1 accounts for 83.58% of the variance, and PC2 accounts for a further 8.17%. Combined, these first two PCs account for over 90% of the total variance and therefore only PC1 and PC2 are required in the analysis.

Figure 8 shows the PCA Scores Plot, for the first two principal components, of the raw Tween20 data.

Figure 8 shows that Thane is the most separate cluster; suggesting that there is a noticeable difference between the final product produced there and the rest of the Croda sites. However, Singapore, Mill Hall and Rawcliffe Bridge are all extremely close together with visibly overlapping data points, which indicates an exceptionally similar final product is produced by these sites. Finally, Chocques appears to be somewhat separate from the main cluster; however, there is still some overlap, and thus there appears to be some similarities between the final product produced there and those produced at Mill Hall, Rawcliffe Bridge and Singapore.

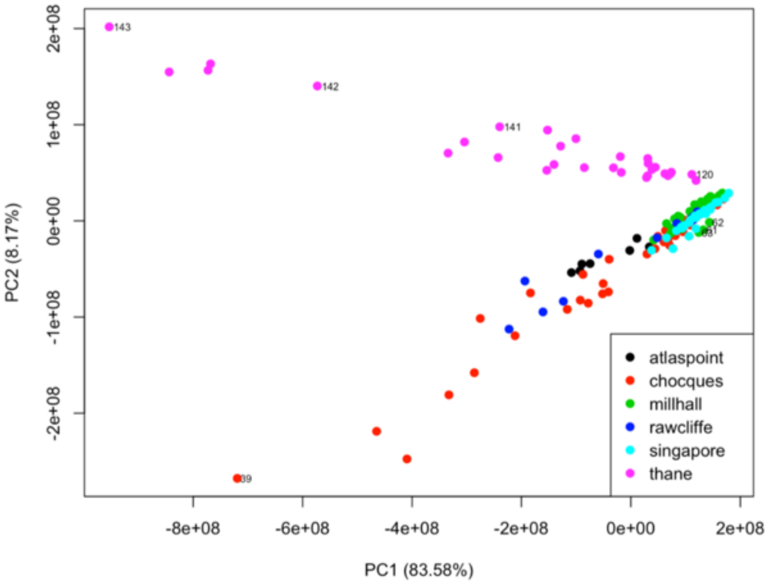


Figure 8. **PCA Scores Plot showing PC1 against PC2 for the raw Tween20 data**. A clear clustering of the data points for Singapore, Mill Hall and Rawcliffe Bridge can be seen. Thane and Chocques data points both exhibit large within group variance with Thane clearly the most separate site cluster. The labelled points 141, 142 and 143 are replicates of the same sample and demonstrate that variance in the MALDI-TOF-MS analysis has an important and undesired impact on PCA. The labelled point 120 shows how much variation there is within the Thane data. Observation 61, 62, 63 appear to differ from the rest of the Mill Hall observations.

Table 2. **Sum of variables for selected observations**. Table 2 demonstrates the variance in the sum of variables (total ion count) for the raw Tween20 data. Clearly, there is a large variation in the total ion count; in particular, replicates (141, 142, 143) of the same observation have notably different total ion counts. As these replicates are of the same observation, they should have the same total ion count and these differences must be due to the experimental method. Therefore, this is a great source of unwanted variation in the analysis. Normalising the data overcomes this by ensuring the sum of variables is the same (10,000) for all observations.

Observation	**Sum of Variables (Total Ion Count) in Billions**
39	21.60
61	2.25
62	1.96
63	2.42
141	10.130
142	16.53
143	23.57

Table 2 shows that the sum of variables (total ion count) for the raw Tween20 data points varies largely. For example, observation 143 has a sum of variables that is twelve times greater than that of observation 62 for the raw data. This large variation in the sum of the variables for the raw data suggests that certain variables may be dominating the analysis for some observations. This could lead the analysis in the wrong direction, likely skewing results and missing out important inferences from underestimated variables. Whereas, the sum of variables for the normalised data is the same for each observation and thus avoids any issues arising from certain overly influential variables distorting the analysis. Moreover, observations 141, 142 and 143 are replicate measurements of the same sample and should therefore have the same total ion count. However, since there is a substantial difference for the total ion count between these observations for the raw data, it is clear that this is a great source of variation in the analysis. Hence, normalising the data results in the total ion count being the same for all three replicates and thus the variance is reduced between them, as desired.

By comparing the PCA scores plot for normalised data (Figure 9) with Figure 8, it is possible to determine whether or not normalisation is better for this analysis.

From Figure 9 it is observed that normalising the data leads to a better visualisation of the difference between sites than the original, non-normalised data (see Figure 8 for comparison). This infers that identifying differences between the sites' final products and their potential causes will be easier and more informative using normalised data. As a result, future analyses will use normalised data.

Other key observations from Figure 9:

- Relatively good separation of data by manufacturing site.
- As with Figure 8, we observe that Thane is the most separate cluster (clearly, the most separate along PC1 and quite separate along PC2) and thus likely producing a final product which is the most dissimilar to the other sites.
- Chocques, Rawcliffe Bridge and Atlas Point demonstrate a clear clustering and overlap of data points, and are therefore likely outputting a near indistinguishable final product.
- Atlas Point and Rawcliffe Bridge appear to be the most similar producers of Tween20. Rawcliffe Bridge and Chocques are then the next two most similar producers, with Atlas Point and Chocques marginally less similar.
- Mill Hall and Singapore are seen to be quite separate from the main cluster of Chocques, Rawcliffe Bridge and Atlas Point. Hence, there could be noticeable dissimilarity between these two sites and the main cluster's final product.
- There appears to be three potential outliers associated with Mill Hall: points 61, 62, and 63. These all come from the same original sample; therefore, it is possible that this particular original sample is an outlier. Before removing any data points as outliers it is important to investigate why they are different and are therefore retained.
- There is a clear pattern followed by all site clusters. Each site cluster follows a near straight-line correlation along PC1 and 2. These patterns or correlation lines likely represent a hidden and non-random variation in the original Tween20 data (the patterns are less clear in Figure 8, but are present for most site clusters there too). It is possible that this hidden variation could be due to the experimental set-up; for example, biases caused by the placement of samples in the matrix. It should be noted that a common misconception that can arise here is that such patterns within clusters implies that PC1 and PC2 are not orthogonal. However, this is not the case since the PC vectors, which are linear combinations of the original variables, are by construction orthogonal to

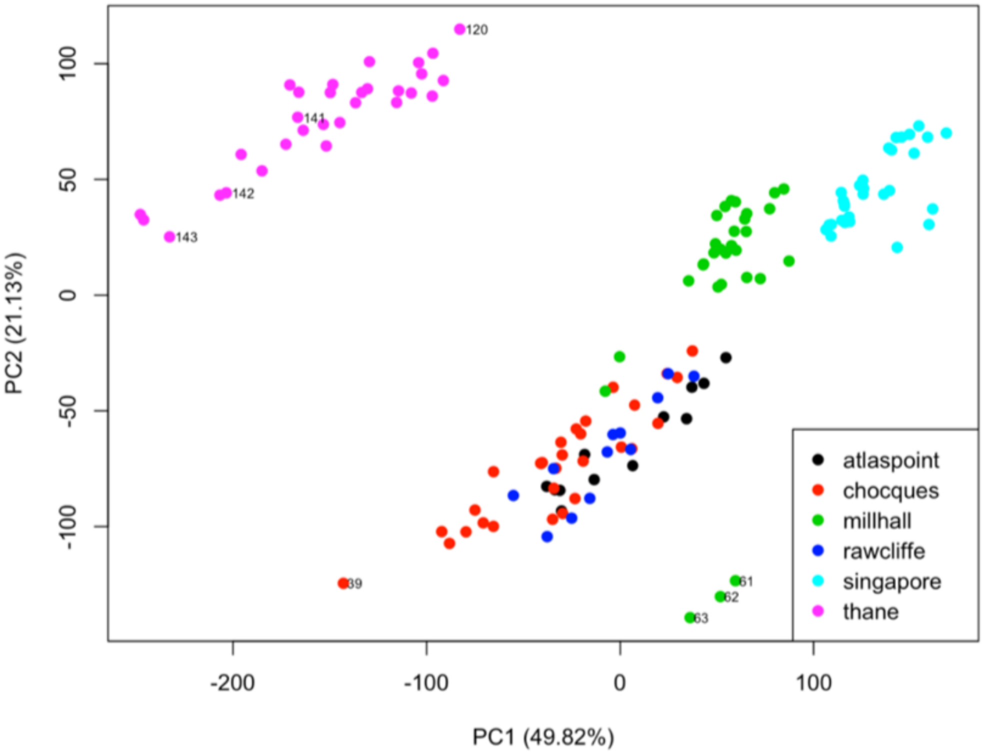


Figure 9. **PCA Scores Plot showing PC1 against PC2 for the Normalised Tween20 data**. Better visualisation of the separation of the data seen here than in Figure 8; which is better for future analyses since identifying differences between sites will be simpler. Generally, there is a clear separation between the manufacturing sites and Thane is seen to be distinctly distant from the other sites; implying the final products produced at Thane will likely be markedly different to those produced at the other sites. There is clear overlap between Atlas Point, Chocques and Rawcliffe Bridge, indicating that the final products at these sites will likely be exceptionally similar. Mill Hall and Singapore appear to produce final products which are somewhat different to the main cluster. Mill Hall does appear to have three potential outliers (61, 62, 63); however, they are not removed since it is important to investigate why they are different before confidently removing them as outliers. Observations 141, 142 and 143 are labelled because they are replicates of the same sample and comparing with Figure 8 the difference between them appears to be less (which is what one would expect as they are the same sample).

each other. PCA scores plots show the magnitude of PC1 against PC2 across the different samples not the orthogonality of the linear combinations of the original variables.

5.2. Squared Mahalanobis Distance. In Table 3, we see the between groups squared Mahalanobis distance, calculated using the first two principal components, for Tween20. This measure takes into account the covariances between observations and, since the PCs are uncorrelated, only the diagonal elements of the covariance matrix (that is, the variances) will be non-zero.

Table 3. Between Sites Mahalanobis Distance with Weighted Covariance for Tween 20.

	Atlas Point	Chocques	Mill Hall	Rawcliffe Bridge	Singapore	Thane
Atlas Point	0	4.79	4.63	0.28	30.36	675.99
Chocques		0	8.10	3.16	33.25	370.07
Mill Hall			0	8.15	17.14	77.30
Rawcliffe Bridge				0	39.65	472.37
Singapore					0	334.37
Thane						0

Key observations from Table 3 are:

- Thane is identified as the site producing the least similar product to the other sites. This agrees with the PCA scores plot in Figure 9. However, the measure appears to misjudge the magnitude of this dissimilarity; that is, it suggests the distance between Thane is far greater than we would expect it to be from what is observed in Figure 9.
- Atlas Point and Rawcliffe Bridge are correctly identified as the most similar locations, producing the most similar final product. This seems quite reasonable since it compares well with Figure 9.
- These results also indicate that Atlas Point and Chocques produce similar final products, in agreement with Figure 9. However, the magnitude of this similarity between these sites according to this metric appears to be somewhat larger than we would have expected from the visual interpretation of the data provided by Figure 9. This is especially obvious when we consider that Atlas Point and Mill Hall are, according to this metric, more similar than Atlas Point and Chocques are, which does not agree with Figure 9. Mill Hall should be a much greater distance from Atlas Point than Chocques; however, this metric has completely failed to recognise this and is therefore inaccurate here.
- Rawcliffe Bridge and Chocques are correctly measured to be the second most closely matched sites by this metric.
- The magnitude of this similarity between Mill Hall and the other sites in particular, appear largely inaccurate when compared to Figure 9. For example, the metric seems to have overestimated the difference between the final products produced by Mill Hall and Singapore and underestimated its difference with Chocques to a considerable degree. Also, the similarity between Mill Hall and Atlas Point appears to be far too small, especially when compared to the similarity between Atlas Point and Chocques (which is greater according to this metric, when in fact the PCA plot clearly shows Atlas Point is substantially more similar to Chocques than Mill Hall). Possibly the potential outliers have caused the issues with Mill Hall since outliers are known to adversely affect the performance of this measure [**31**]. For this reason, outliers must be investigated further.

Although this metric does appear to be somewhat successful when comparing the more closely matched sites (Atlas Point, Chocques, and Rawcliffe Bridge), results in Table 3 fail to fully reflect Figure 9, indicating that the measure is not reliable for this data set. Other similarity measures must be experimented with for comparison.

5.3. Bhattacharyya Distance. In Table 4, the between groups squared Bhattacharyya distance, calculated using the first two principal components, for Tween20 is provided.

TABLE 4. Between Sites Bhattacharyya Distance with Weighted Covariance for Tween20.

	Atlas Point	Chocques	Mill Hall	Rawcliffe Bridge	Singapore	Thane
Atlas Point	0	0.70	1.26	0.09	3.99	84.65
Chocques		0	1.48	0.48	4.30	46.31
Mill Hall			0	1.49	2.52	10.35
Rawcliffe Bridge				0	5.06	59.25
Singapore					0	42.01
Thane						0

Important remarks from Table 4 are:

- Thane is identified as the most separate site cluster and Singapore as the second most (as was the case with the Mahalanobis equivalent). This agrees with the visualisation in the PCA scores plot shown in Figure 9.
- The overall order of similarity (Rawcliffe Bridge and Atlas have the smallest non-zero distance and are most similar, then it is Atlas Point and Chocques, and so on) appears to mostly reflect what is observed in Figure 9; however, it is clear that this measure is vastly overestimating the differences between the sites. This is similar to how the Mahalanobis measure performed.
- The measure correctly identifies Atlas Point and Rawcliffe Bridge as the most similar sites and appears quite accurate compared with Figure 9 for the relationship between Atlas Points, Chocques and Rawcliffe Bridge sites.

Overall, this metric offers a more reliable measure than the Mahalanobis distance. For example, it correctly identifies that Chocques is more similar to Atlas Point than it is to Mill Hall, whereas the Mahalanobis distance did not. However, it is still inaccurate; especially, in terms of the magnitude of dissimilarity between the Croda sites. It is also notably unsuccessful at accurately measuring the difference between the most separate site clusters and the rest. That is, the distances in Table 4 for Thane, for example, are far larger than one would expect from the visualisation of the data shown in Figure 9. Therefore, this measure is not practical to use.

REMARK. The Hellinger distance was also considered; however, it readily become apparent during initial findings that this measure merited no further investigation or discussion.

5.4. Within Group Variance, Between Group Variance, and Separation. To account for the percentage of the variance in the Tween20 data due to each of the first two PCs, α values are used.

We can simply calculate the α values as:

$$\alpha_i = \frac{var_{PCi}}{var_{PCi}+var_{PCj}} \text{ for } i \neq j, i = 1,2 \text{ and } j = 1,2.$$

Where var_{PCi} and var_{PCj} are the percentage of the variance in the analysis along PCi and PCj, respectively.

Hence, for the first two PCs we have:

$$\alpha_1 = \frac{49.73}{49.73+20.43} \approx 0.7, \alpha_2 = \frac{20.43}{20.43+49.73} \approx 0.3$$

This gives greater weighting to PC1 as desired.

Using the concept of covariance and linear combinations we can apply the following formulae for within and between groups variances (an extension of the formulae discussed in section 4.6),
$Var_{WG}(\sum_{i=1}^{n} \alpha_i X_i) = \sum_{i,j} \alpha_i \alpha_j cov_{WG}(X_i X_j)$ and $Var_{BG}(\sum_{i=1}^{n} \alpha_i X_i) = \sum_{i,j} \alpha_i \alpha_j cov_{BG}(X_i X_j)$
where $cov_{WG}(X_i X_j) = \frac{\sum_{k=1}^{g} (n_k - 1) S^2_{i,j,k}}{\sum_i n_i - g}$ and $cov_{BG}(X_i X_j) = \frac{\sum_{k=1}^{g} n_k (\overline{x}_{i,k} - \overline{\overline{x}}_i)(\vec{x}_{i,k} - \overline{\overline{x}}_j)}{\sum_i n_i - g}$, to the Tween20 data to obtain the following tables.

Given in Tables 5, 6 and 7 are the weighted Within Groups Variance, weighted Between Groups Variance and the Separation, respectively, for Tween20 data. Here, Separation is a natural similarity measure since it provides a means of measuring the similarity between sites. It does this by accounting for both the internal (within groups) variance and external (between groups) variance. The smaller the Separation, the greater the similarity.

Table 5. Within Groups Variance for Tween20 Weighted for PCs 1 and 2.

	Atlas Point	Chocques	Mill Hall	Rawcliffe Bridge	Singapore	Thane
Atlas Point	63731.15	72614.20	50692.93	54192.74	29539.73	86234.89
Chocques		75983.64	60865.52	67368.08	46277.10	85377.22
Mill Hall			45747.40	45446.81	31158.99	70259.10
Rawcliffe Bridge				44654.33	24293.61	80988.77
Singapore					16570.57	55670.68
Thane						94770.79

Between Group Variance for Tween20 Weighted for PCs 1 and 2 is given in Table 6.

Table 6. Between-Groups Variance for Tween20 Weighted for PCs 1 and 2.

	Atlas Point	Chocques	Mill Hall	Rawcliffe Bridge	Singapore	Thane
Atlas Point	0	10063.55	74429.83	3815.58	322978.90	165482.10
Chocques		0	168036.37	2076.05	539561.80	113362.10
Mill Hall			0	107303.36	105382.20	557434.30
Rawcliffe Bridge				0	388211.20	123933.10
Singapore					0	1147558.00
Thane						0

Separation for Tween20 Weighted for PCs 1 and 2 are given in Table 7.

Table 7. Separation for Tween20 Weighted for PCs 1 and 2.

	Atlas Point	Chocques	Mill Hall	Rawcliffe Bridge	Singapore	Thane
Atlas Point	0	0.14	1.47	0.07	10.93	1.92
Chocques		0	2.76	0.03	11.66	1.33
Mill Hall			0	2.36	3.38	7.93
Rawcliffe Bridge				0	15.98	1.53
Singapore					0	20.61
Thane						0

Tables 5,6 and 7 show that both the between and within groups variances are somewhat accurate; however, it appears that separation is not useful as a similarity measure for this data. A different approach is needed.

5.5. The Hierarchical Clustering Approach. Hierarchical clustering can be used to derive a similarity measure in a number of different ways. Although not strictly a metric, the best approach found was to use an asymmetric distance. This distance is informative since it describes the interchangeability of sites in the sense that Site A can replace Site B fully, but Site B may not be able to replace Site A. For example, we can examine Table 8, which shows the Asymmetric Similarities for the raw Tween20 data. The bottom row shows how dispersed the data is within a site using a symmetric distance since a greater understanding of within site similarity is achieved. Here, greater than one (> 1) means it is less dispersed than average, less than one (< 1) means more dispersed than average. It is important to take into consideration that this measure is affected by variance within a site.

The asymmetric similarity is calculated by first finding the maximum cophenetic distance between observations within group i and then the maximum cophenetic distance between observations within groups i and j; then, dividing the maximum cophenetic distance between observations within groups i and j by the maximum cophenetic distance between observations within group i and invert to obtain the similarities. The process is the same for symmetric similarity except that the mean of the maximum cophenetic distances within each group is used as the divisor instead (the code for both the asymmetric and symmetric similarities are given in appendix A).

Table 8. Table of Asymmetric Similarities for Raw Tween20 Data.

	Atlas Point	Chocques	Mill Hall	Rawcliffe Bridge	Singapore	Thane
Atlas Point	1	0.70	0.49	1	0.58	0.44
Chocques	1	1	0.71	1	0.84	0.63
Mill Hall	1	1	1	1	1	0.9
Rawcliffe Bridge	1	0.70	0.49	1	0.58	0.44
Singapore	0.55	0.55	0.46	0.55	1	0.41
Thane	0.53	0.53	0.53	0.53	0.53	1
	1.27	0.88	0.63	1.27	1.35	1.05

Inferences from Table 8:

- All sites have 100% similarity with themselves, so are assumed to be internally consistent by the asymmetric measure. Which is not unreasonable.

- CH to AP, RB to AP, CH to RB, and MH to SI are the joint most similar between sites (100%). However, MH to SI appears to be very different to what we would expect, whereas the others do not.
- Sites which are very similar but not 100% according to this measure are: RB to CH (85%), MH to AP (84%), MH to CH (84%), MH to RB (84%), AP to RB (78%).
- Least similar sites: SI to TH (33%), AP to TH (34%), SI to AP (41%), SI to CH (41%), SI to RB (41%), AP to MH (42%), AP to SI (42%), RB to TH (44%), SI to MH (49%), CH to TH (51%).
- MH's low value in the bottom row suggests outliers could be present.

Table 8 can be compared with its corresponding dendrogram shown in Figure 10. Figure 10 provides a visualisation of what is observed from Table 8. AP, CH and RB all appear to produce highly similar final products which would be interchangeable between the sites. The separation between the rest of the site clusters is also clear to see here, matching what is described in the inferences from Table 8. The potential presence of one or more outliers in MH cluster (observed from Table 8) is shown quite clearly in Figure 10; observation 21 is clearly different to the rest of the cluster (noticeable since the cophenetic distance is clearly much greater between 21 and the other MH observations) and is likely to be an outlier.

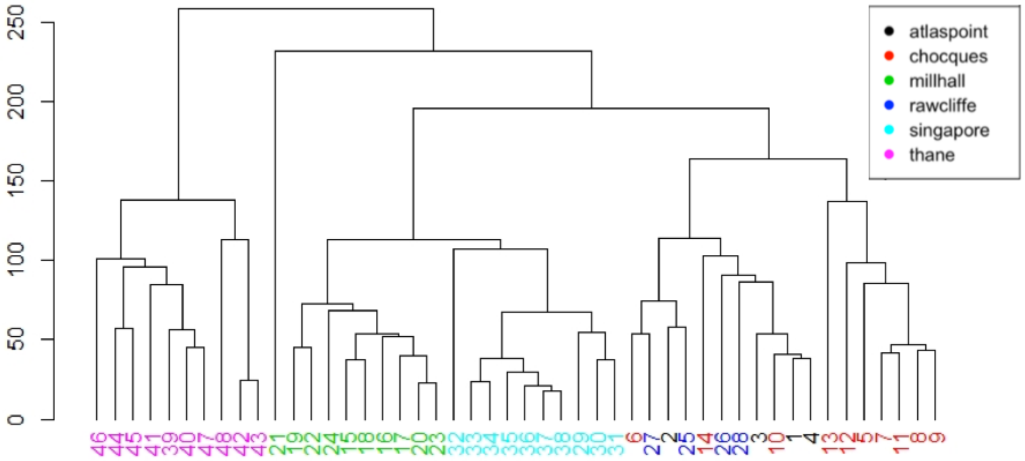


FIGURE 10. **Dendrogram using Euclidean Distance and Average Linkage for Raw Tween20 data.**

To investigate the viability of extracted features (as described in methods, section 4.9, these are FWHM, Maximum Height, Ratio of Heights and Centre of the Gaussian model) we examine Figures 11 and 12 which show the dendrograms for raw and extracted features BrijCS20 data respectively. For ease of comparison, both dendrograms use Euclidean Distance and Average Linkage. Clearly, these Figures indicate that it is possible to use extracted features for this analysis (this is also important for distribution modelling as will be seen in section 5.6).

Tables 9 and 10 show the Asymmetric Similarities for Raw BrijCS20 Data and for Extracted Features BrijCS20 Data, respectively. Like Figures 11 and 12 these tables indicate that it is possible to use extracted features for this analysis since they produce similar results. However, outliers (see observation 2 in Figure 12, for example) are a concern for extracted features that will be investigated.

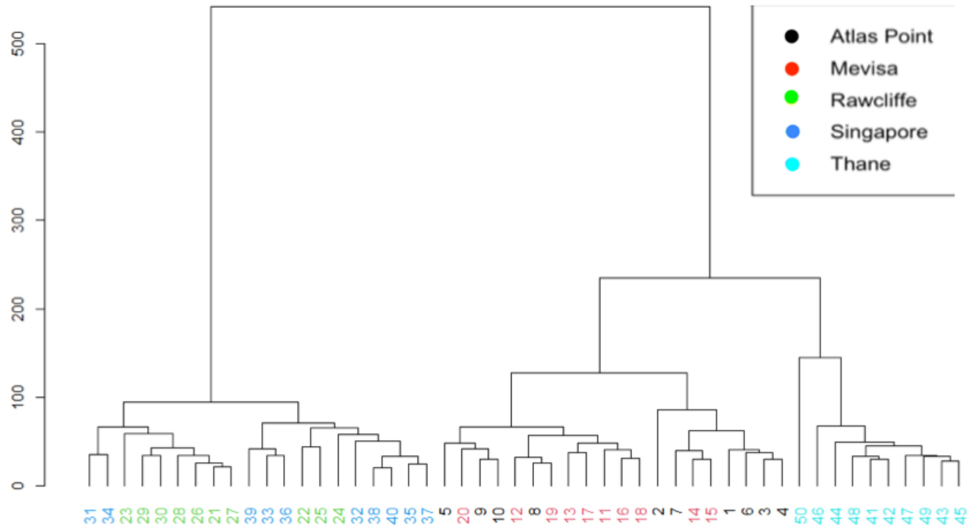


FIGURE 11. **Dendrogram using Euclidean Distance and Average Linkage for Raw BrijCS20 data.**

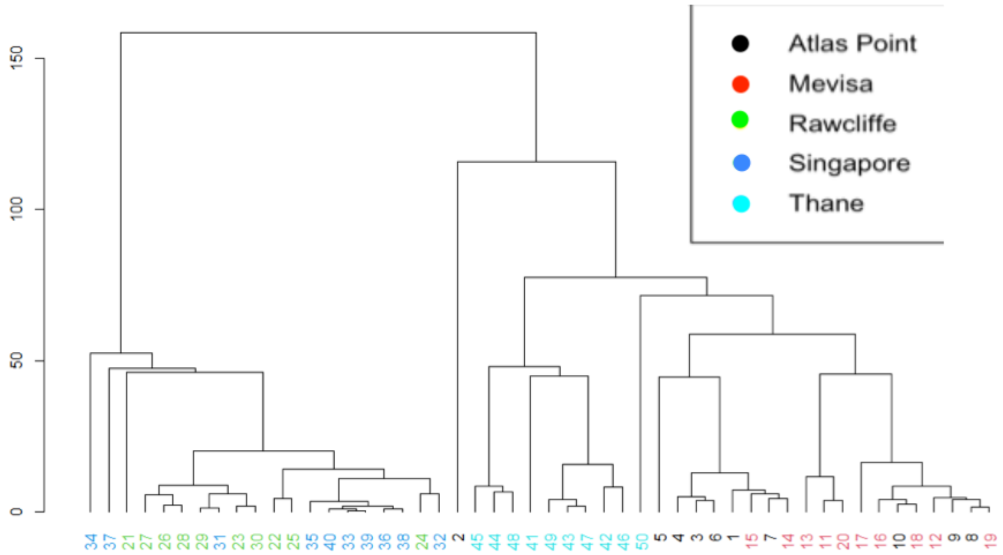


FIGURE 12. **Dendrogram using Euclidean Distance and Average Linkage for Extracted Features BrijCS20 data.**

Table 9. Table of Asymmetric Similarities for Raw BrijCS20 Data.

	Atlas Point	Mevisa	Rawcliffe Bridge	Singapore	Thane
Atlas Point	1	1	0.24	0.24	0.55
Mevisa	1	1	0.24	0.24	0.55
Rawcliffe Bridge	0.18	0.18	1	1	0.18
Singapore	0.18	0.18	1	1	0.18
Thane	0.62	0.62	0.27	0.27	1
	0.92	0.92	1.25	1.25	0.82

Table 10. Table of Asymmetric Similarities for Extracted Features BrijCS20 Data.

	Atlas Point	Mevisa	Rawcliffe Bridge	Singapore	Thane
Atlas Point	1	0.92	0.22	0.22	0.50
Mevisa	0.92	1	0.22	0.22	0.50
Rawcliffe Bridge	0.22	0.22	1	1.25	0.22
Singapore	0.22	0.22	0.25	1	0.22
Thane	0.50	0.50	0.22	0.22	1
	0.61	1.19	1.52	1.34	0.91

5.5.1. *Identifying Outliers.* Univariate boxplots using PC1 scores as input is the first method investigated for detecting outliers. The whiskers for these boxplots are calculated as $w_1 = q_1 - 1.5(q_3 - q_1)$ and $w_3 = q_3 + 1.5(q_3 - q_1)$, where q_1 is the first quartile of the data and q_3 is the third quartile. Hence, $q_3 - q_1$ is the interquartile range. w_1 and w_3 are shown as the long horizontal lines in the boxplots and any data points outside these limits are defined to be outliers. Figure 13 shows the univariate boxplot for the Mill Hall data from Tween20.

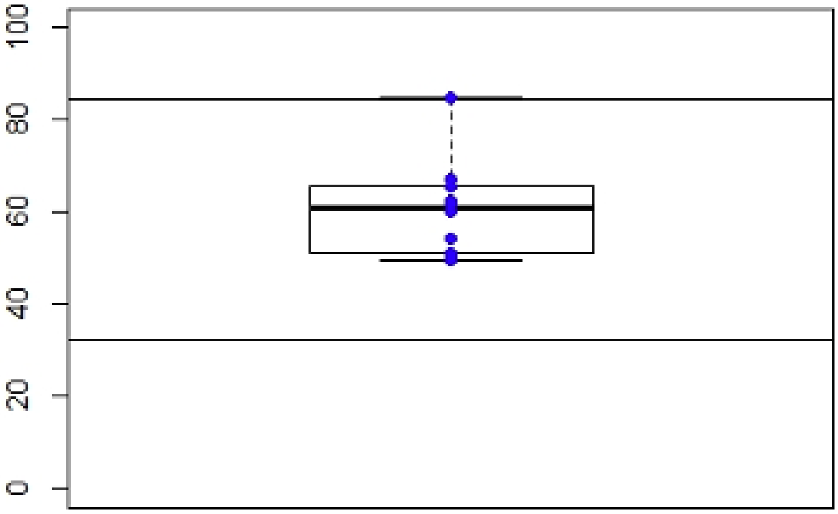


Figure 13. **Univariate Boxplot for Mill Hall Data from Tween20**

This boxplot (Figure 13) indicates that observation 16 is an outlier, which does not make sense and is not reflective of what is seen in the PCA scores plot or the dendrogram. This univariate boxplot method produced reasonable results for the other sites; however, the Mill Hall data is crucial since it contains the outlier observed in the dendrogram. Since this boxplot does not correctly identify the outlier and, in fact, misclassifies observation 16 as such, it is concluded that this method is ineffective.

The second approach tried in outlier detection uses the first two PCs with the Mahalanobis Distance. Figure 14 shows the boxplot produced under this method for Mill Hall and Figure 15 shows the boxplot for the Singapore data.

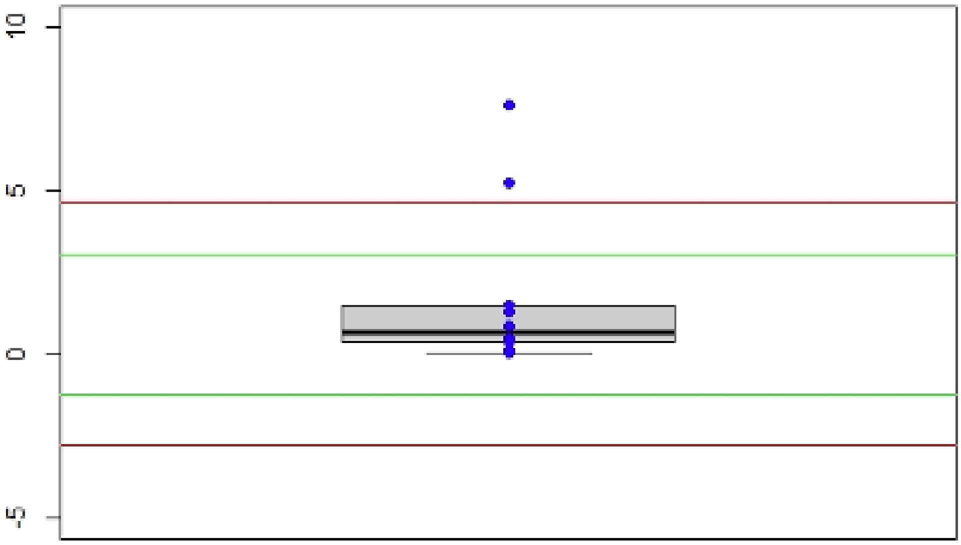


Figure 14. **Mahalanobis Boxplot for Mill Hall Data from Tween20**

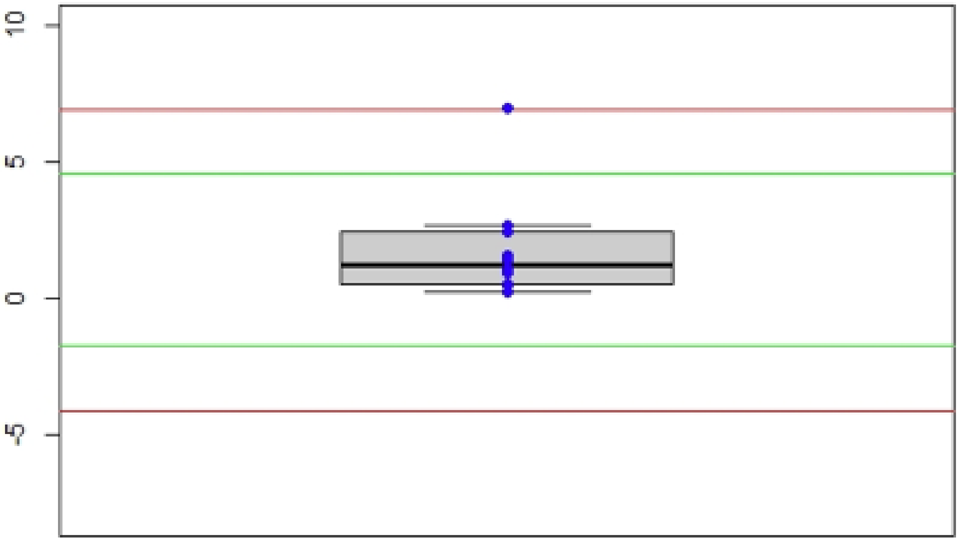


Figure 15. **Mahalanobis Boxplot for Singapore Data from Tween20**

As Figures 14 and 15 show, this approach is close. It correctly identifies observation 21 as an outlier for Mill Hall, but unfortunately observation 16 from Mill Hall and observation 32 from Singapore are misidentified as outliers too. This means that this method is likely too sensitive.

A novel approach to this problem was to examine the use of bivariate Bagplots using PC1 and PC2 scores as input. Figure 16 gives the bagplots produced for each site of the Tween20 data.

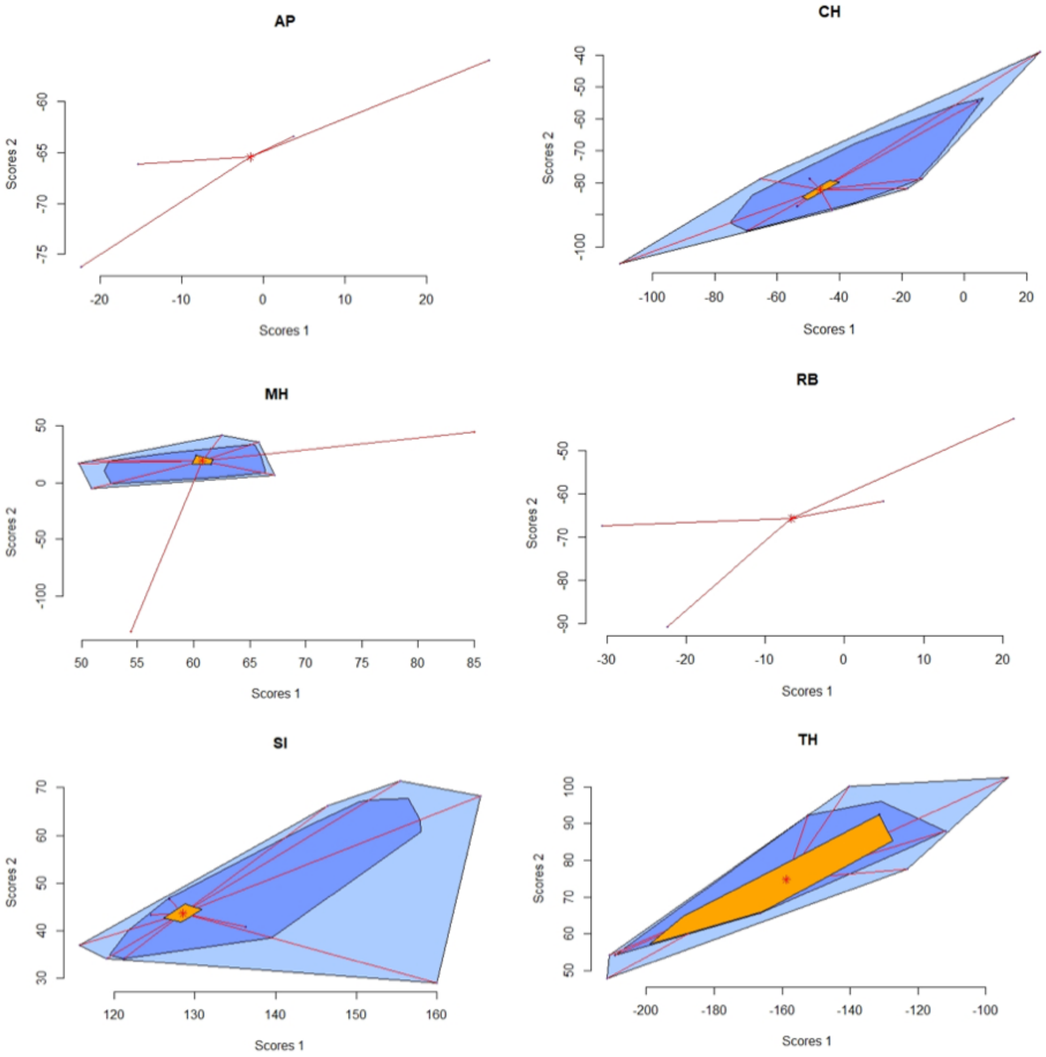


Figure 16. **Bagplots for Tween20 data by site. From top left to bottom right: Atlas Point, Chocques, Mill Hall, Rawcliffe Bridge, Singapore, Thane.**

Though somewhat successful for larger groups of data this method (Figure 16) ultimately fails as it does not perform adequately for the small data subsets; in particular, it fails to work for Rawcliffe Bridge and Atlas point data (both only having four data points each).

A fourth approach uses decision boundaries or fences to discriminate between outliers and valid data. This method involves using PC1 and PC2 scores as input and constructs the fences as follows:

Inner fences: $w_1 = q_1 - 1.5(q_3 - q_1)$ and $w_3 = q_3 + 1.5(q_3 - q_1)$.
Outer fences: $w_1 = q_1 - 3(q_3 - q_1)$ and $w_3 = q_3 + 3(q_3 - q_1)$
Where all variables are as defined in the univariate boxplot method.

Figures 17 and 18 demonstrate how this method performs for Chocques and Mill Hall, respectively.

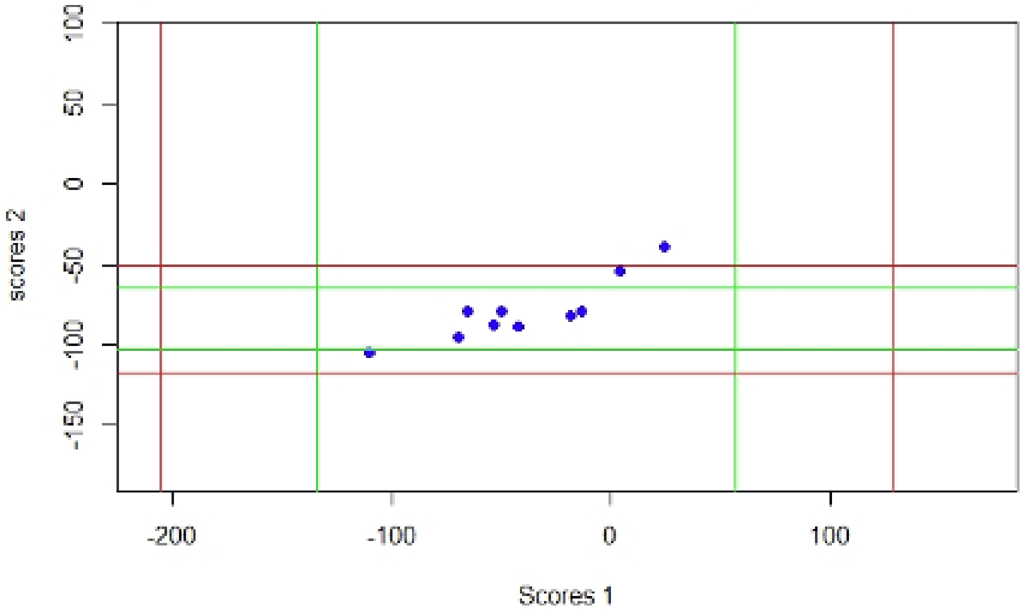


FIGURE 17. **Decision Boundary Boxplot for Chocques Data from Tween20**

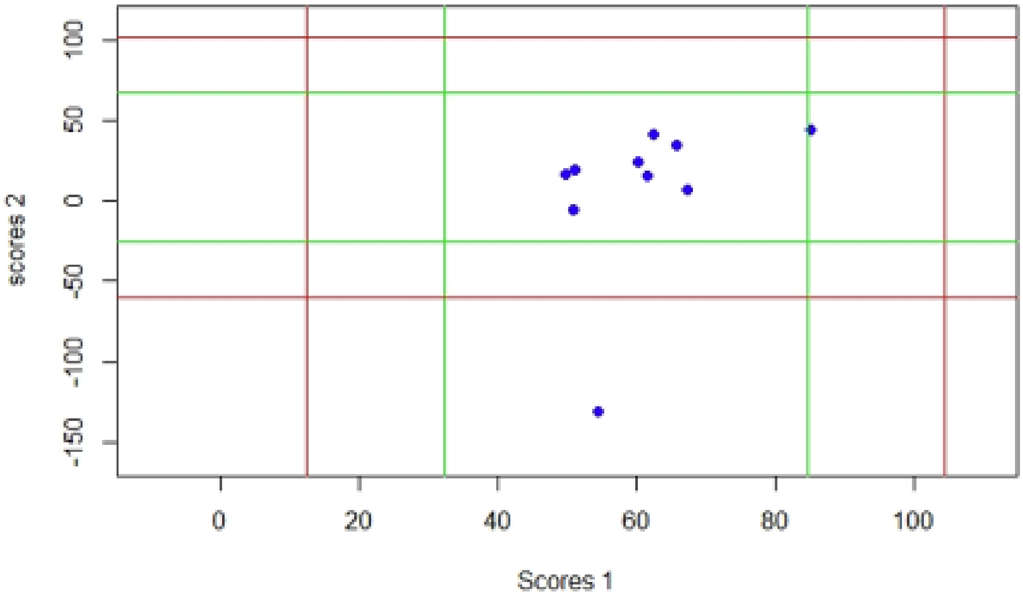


FIGURE 18. **Decision Boundary Boxplot for Mill Hall Data from Tween20**

Clearly, an outlier is found where there should be none for Chocques (Figure 17), meaning that this method is overly sensitive. It does perform well for Mill Hall, finding observation 21 to be an outlier as we would have expected. This is yet another approach which came close; however, it is ultimately still unsuccessful.

Unfortunately, after numerous ideas and attempts, no perfect solution was found for this crucial issue. All four methods discussed here come close; however, the lack of data for some of the manufacturing sites is problematic. With more data in the future, this problem may be solved.

5.6. Distribution Modelling.

5.6.1. *Gaussian Model.* All parameter values (μ and σ) are calculated as described in Methods and then stored in a matrix from which the model Gaussian curves are fitted to each distribution. A selection of these fitted curves are shown in Figure 19, to illustrate how well the fitted model works for BrijCS20 data. Here, the two largest and therefore important distributions are plotted with overlaying Gaussian fits along their peaks. These distributions are related peaks from the same compound (the BrijCS20 product). They differ in shape due to the number of ethylene oxide molecules (also known as EO units) added to them. To identify the distributions we can use the consistent difference of 44 Daltons, which correspond to the EO units, in their m/z values.

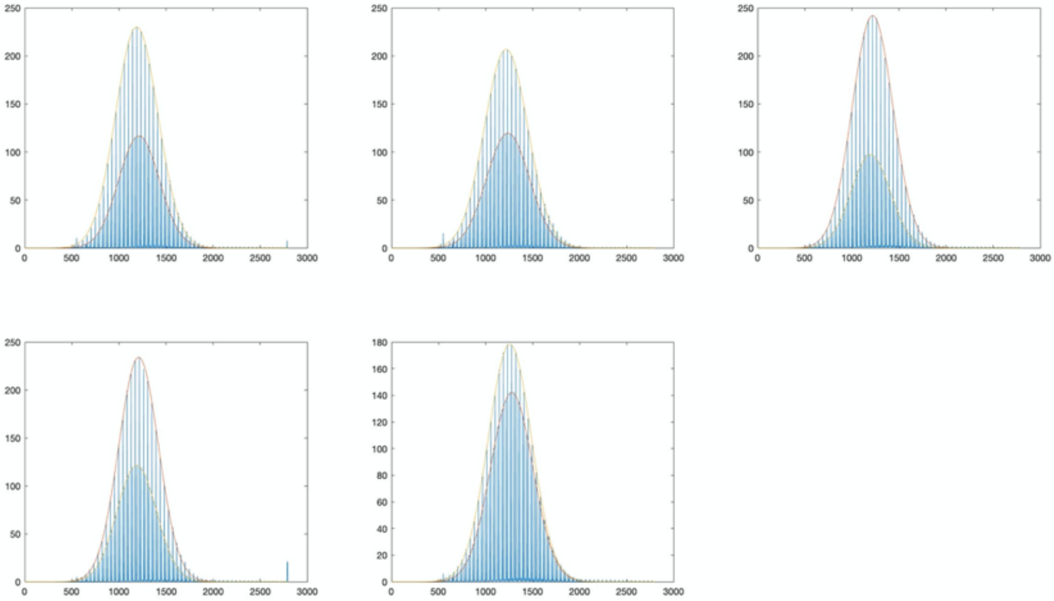


Figure 19. **Gaussian Fits for BrijCS20 data**. From top left to bottom right: Observation from Atlas Point, Observation from Mevisa, Observation from Rawcliffe, Observation from Singapore, Observation from Thane. Plot of the data for observations from the different sites with the fitted Gaussian curves overlaid. Two largest distributions in the BrijCS20 data are fitted with Gaussian curves. The curves appear to fit the data reasonably well in each case; therefore, it is possible to cautiously assume the data is Gaussian. There is some error, especially for the higher masses, which may indicate that there could be some right skew in the distribution.

Figure 19 demonstrates that the Gaussian models fit the data quite well. Thus, we can cautiously use the assumption that the data is Gaussian for these two selected distributions. This means that we can use properties of the Gaussian distribution to analyse the data. For example, we extract the maximum height, FWHM, the ratio between the two distributions‘ maximum heights and the centre of the distribution, and use them as variables in further analyses.

To verify how accurate the Gaussian fit is for this data, the difference between the actual (experimental) data and calculated or modelled Gaussian data is calculated. An example of these calculated errors are provided in Table 11, which includes a sample of the calculated errors for the first distribution for observation 1 from Atlas Point (plot is given in Figure 19).

Table 11. **An Example of calculated errors for the Gaussian Fit.** Table 11 shows a selection of the errors for the Gaussian fit. Observation 1 from Atlas Point is used in this example, and, although less accurate for larger masses, it demonstrates that the fitted model appears reasonably accurate for this data.

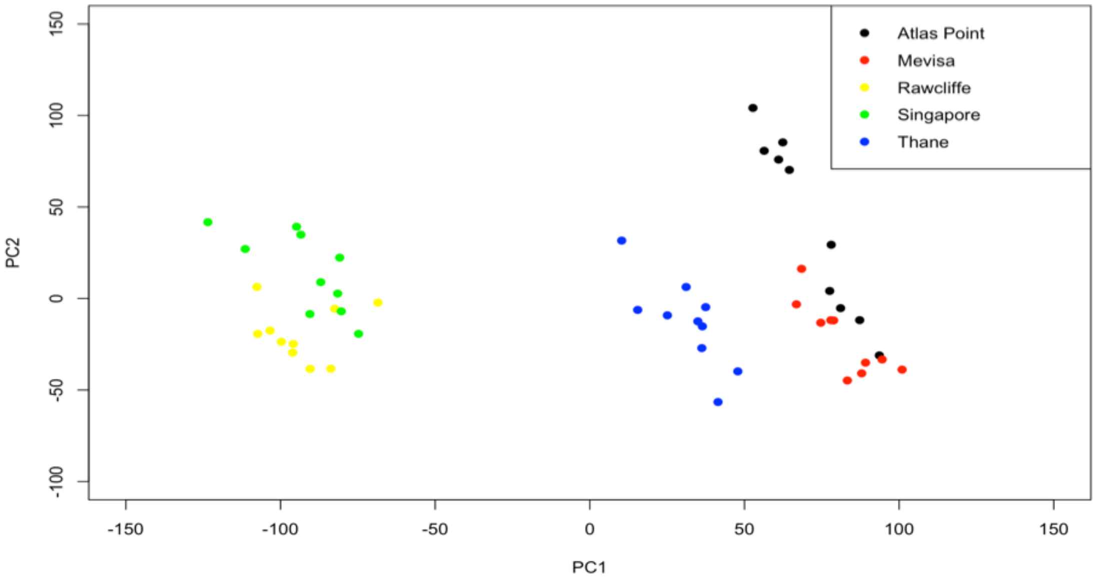
m/z value	Error between Gaussian Fit and Actual Data
882	0.48
1014	0.91
1146	0.57
1278	0.32
1410	0.64
1542	1.98
1674	2.72
1806	2.46

Table 11 demonstrates a common trend for the Gaussian fit errors; namely, that for higher masses the accuracy of the model decreases. The slight decrease of accuracy in the model for these higher masses is not substantial and is likely to be attributed to the greater variance often observed for large masses in mass spectrometry. This fact, coupled with the small error observed between the fitted Gaussian distribution and the actual data for all masses, indicates that the data can be modelled as Gaussian.

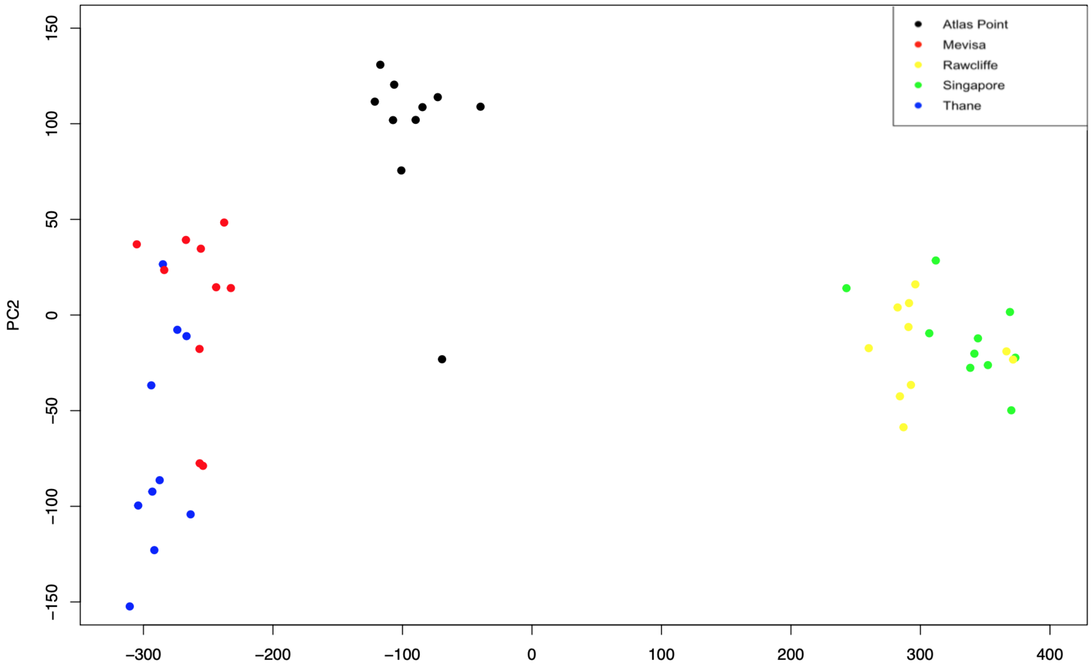
Continuing under the assumption that the Gaussian model is suitable, the PCA scores plot for the extracted variables (FWHM, MZ centre, distribution ratio and maximum intensity) for BrijCS20 data is provided in Figure 20(A). 20(B) shows the PCA scores plot for the original normalised Brijcs20 data and is used for comparison with 20(A).

The most important observation from Figure 20(A) is that it closely resembles Figure 20(B). This key insight tells us that the extracted variables provide a similar overall representation of the data as the original data; meaning that the analysis can be simplified in this way without a significant loss of information from the data. Since the same overall pattern is observed in both Figure 20(B) and 20(A), the key observations are similar. That is, the separation observed between the different clusters is extremely similar in 20(B) and 20(A). Arguably, the potential Thane outlier appears somewhat less prominent in Figure 20(A); however, this is marginal and thus not a concern.

When used in similarity measures such as Bhattacharyya and Mahalanobis the extracted features data and Gaussian model failed, this is likely due to the issues caused



(A) PCA Scores Plot for Extracted Features BrijCS20 data.



(B) PCA Scores Plot for Original BrijCS20 data.

FIGURE 20. **Showing the PCA Scores Plot for Extracted Features BrijCS20 data (A) and the PCA Scores Plot for Original BrijCS20 data (B) for comparison (that is, to see whether extracted features can be used).**

by outliers (see section 5.5.1). Other distributions including Fréchet and Weibull were also experimented with, but they failed to fit the data any better than the Gaussian model.

For additional insight and analysis we also investigate the Glycerox HE product. In Figure 21 the peaks of the distributions are joined to form the distributions of the samples for Atlas point. Clearly, the data points (observations 1 to 9 from the Glycerox data) do not follow a normal distribution fully since there is some right-skewness for all samples. This again highlights that the data, although close, does not fully follow the normal distribution and as such extracted features from the distribution may not be as useful as hoped. The data is far more skewed for Glycerox and this is likely due to the products composition.

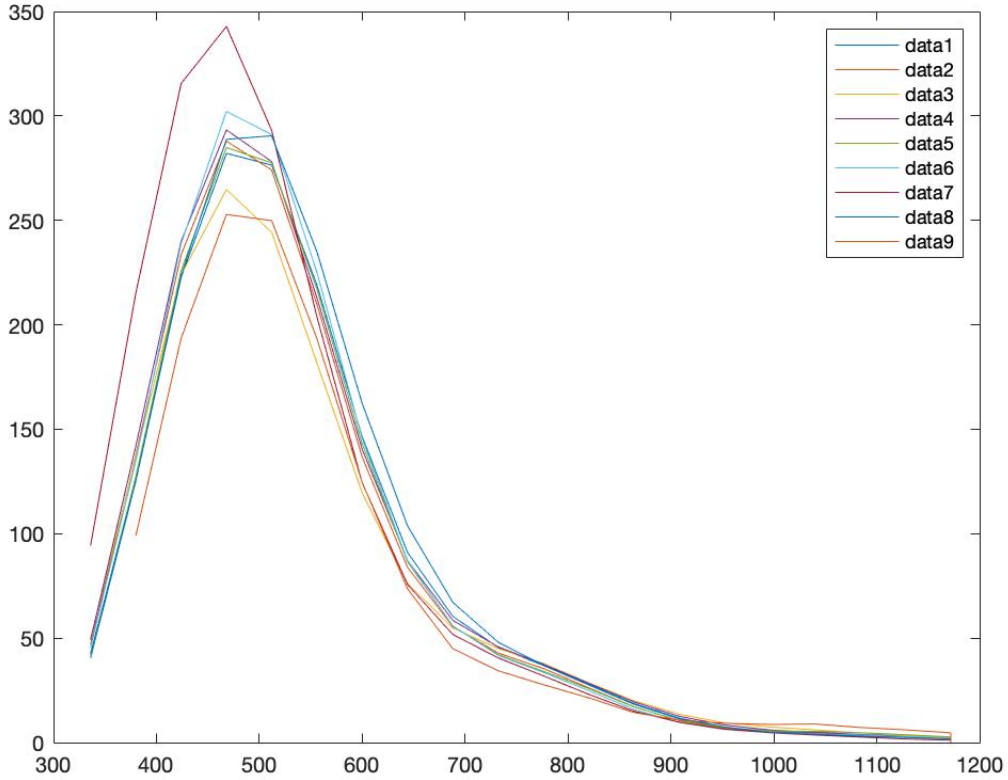


Figure 21. **Figure 21, showing the peaks of the distributions using joined lines for each sample of Glycerox from Atlas Point.** A clear right-skew is seen in the distribution, indicating that the Glycerox HE data does not follow the Gaussian distribution.

Future analyses could look at other novel distributions or combinations of distributions to produce a more viable and practical model.

6. Conclusion

This research showed that finding a suitable similarity measure for Croda's chemical products is challenging. However, numerous approaches yielded promise, and it is probable that a practical solution can be found. Distribution fitting, as well as the hierarchical clustering approach, are particularly encouraging. The Gaussian curve nearly fits the data and with further investigation another distribution or combination of distributions that better fits the data may be found. Similar promise is shown with the asymmetric similarity measure from hierarchical clustering, especially with regard to the accuracy of internal consistency of products. However, the issues with outliers need to be resolved for both of these approaches. If they are, then either of these methods may provide a practical similarity measure for Croda. It is also clear that more data would help substantially in this research, especially for developing a means for identifying outliers. Perhaps, bagplots or a variation of the boxplots tried may then result in an effective outlier detection algorithm.

Other approaches, investigated in the early stages, did not produce much promise. The Bhattacharyya and Mahalanobis distances seemed, intuitively, to be prime candidates as measures for this data; however, the results were unexpectedly worse than hoped. These measures did not align well with the data and the separation observed in principal component analysis. Separation was investigated as the next best candidate. Again, the results were not accurate enough and separation, despite the within and between groups variances showing reasonable accuracy, was dismissed as an option for Croda's similarity measure. More novel and complex approaches were then investigated, these include distribution fitting and the hierarchical clustering approaches discussed above.

Overall, no approach investigated in this project produced the similarity measure, the simple decision boundary, from which Croda could confidently and easily determine whether or not a sample of a product was consistent enough to use. Yet, this research is not in vain since avenues for further progress have appeared with both distribution fitting and hierarchical clustering. One such avenue is to model the data with another distribution or a combination of distributions and use the properties of that distribution to find a more suitable and informative measure.

Bibliography

[1] About us | Croda, Croda PLC, viewed 15 October 2020, https://www.croda.com/en-gb/about-us.

[2] Our History | Croda, viewed 15 October 2020, https://www.croda.com/en-gb/about-us/our-history.

[3] Land | Croda, Croda PLC, viewed 16 October 2020, https://www.croda.com/en-gb/sustainability/our-sustainability-in-action/land#tab-collapse-domestic-material-consumption.

[4] Kolossváry, I. and Wegscheider, W. (1990), A similarity measure for chemical data: Applications to cluster analysis. J. Chemometrics, 4. pp. 255-266.

[5] Kowalski, B. R. and Bender, C. F. (1972), Pattern recognition. Powerful approach to interpreting chemical data. Journal of the American Chemical Society Vol. 94 Issue 16. pp. 5632 - 5639.

[6] Tetko, I.V., Engkvist, O., Koch, U., Reymond, J.L. and Chen, H., 2016. BIGCHEM: challenges and opportunities for big data analysis in chemistry. Molecular informatics, 35(11-12), pp.615-621.

[7] Tetko, I.V. and Engkvist, O., 2020. From Big Data to Artificial Intelligence: chemoinformatics meets new challenges.

[8] Ayorinde, F.O., Gelain, S.V., Johnson Jr, J.H. and Wan, L.W., 2000. Analysis of some commercial polysorbate formulations using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Communications in Mass Spectrometry, 14(22), pp.2116-2124.

[9] Tween20 | Croda Personal Care, Croda PLC, viewed 15 October 2020, https://www.crodapersonalcare.com/en-gb/products-and-applications/product-finder/product/2116/Tween_1_20.

[10] Pavoni, L., Perinelli, D.R., Ciacciarelli, A., Quassinti, L., Bramucci, M., Miano, A., Casettari, L., Cespi, M., Bonacucina, G. and Palmieri, G.F., 2020. Properties and stability of nanoemulsions: How relevant is the type of surfactant? Journal of Drug Delivery Science and Technology, 58, p.101772.

[11] BrijC20 | Croda Personal Care, Croda PLC, viewed 14 January 2021, https://www.crodapersonalcare.com/en-gb/products-and-applications/product-finder/product/1939/Brij_1_CS20

[12] Glycerox HE | Croda Personal Care, Croda PLC, viewed 28 May 2021, https://www.crodapersonalcare.com/en-gb/products-and-applications/product-finder/product/306/Glycerox_1_HE.

[13] Atkins, P. and De Paula, J. (2006), Atkins' Physical Chemistry Eighth Edition. Oxford University Press. pp. 655-656.

[14] Fréchet M. Rendic. Circ. Mat. Palermo 22 (1906) 1-74.

[15] Pearson, K. 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, Series 6, 2(11), 559-572.

[16] Webb, A. R. (2003). Statistical Pattern Recognition. John Wiley and Sons. pg. 321-324.

[17] Hotelling, H. 1933. Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology, 24(6 & 7), 417-441 & 498-520

[18] Borgatti, S.P. Handout Normalising Variables, viewed 25 November 2020. www.analytictech.com/ba762/handouts/normalization.htm

[19] McLachlan, GJ. (1999), Mahalanobis Distance. Resonance. Springer.

[20] Deza E. and Deza M.M. (2006), Dictionary of Distances, Elsevier. pg. 241-254.

[21] Wehrens, R., Hageman, J.A., van Eeuwijk, F. et al. (2016). Improved batch correction in untargeted MS-based metabolomics. Metabolomics 12, 88.

[22] Tilevik, A., 2017. Evaluation of clustering methods for analyzing drug cytokine profiles. viewed 14 June 2021, http://urn.kb.se/resolve?urn=urn:nbn:se:kau:diva-64242.

[23] Ward J.H., Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association. 1963, 58:236-244.

[24] Webb, A. R. (2003). Statistical Pattern Recognition. John Wiley and Sons. pg. 363-371.

[25] J.W. Tukey. Exploratory data analysis. Addison-Wesley, Reading, 1977.

[26] Dekking, F.M., Kraaikamp, C., Lopuhaa, H.P., Meester, L.E.(2005) A Modern Introduction to Probability and Statistics Understanding Why and How, Springer. Pg 237.

[27] Rousseeuw, Peter J.; Ruts I.; Tukey J. W. (1999). “The Bagplot: A Bivariate Boxplot”. The American Statistician. 53 (4): 382-387.

[28] Bagplot, Wikipedia.org, viewed 25 January 2021, https://en.wikipedia.org/wiki/Bagplot.

[29] C.F. Gauss. Theoria motus corporum coelestium in sectionis conicis solem am-bientum. In: Werke. Band VII. Georg Olms Verlag, Hildesheim, 1973. Reprint of the 1906 original.

[30] N. Ernst, G. Bozdech, H. Schmidt, W.A. Schmidt, Grover L. Larkins, On the full-width-at-half maximum of field ion energy distributions, Applied Surface Science, Volume 67, Issues 1-4, 1993, Pages 111-117.

[31] Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015), A comparison study on similarity and dissimilarity measures in clustering continuous data. PloS one, 10(12), e0144059.

Appendices

Since there is a vast amount of extremely similar code in this project, the appendices are used wisely, containing the key segments of code that represents the work in this project.

A. R Code for Cophenetic Distances and Derived Similarity Measure

```
# Script for the asymmetric version coph = function (data, groups, ngroups)
{
var < – matrix(0, ncol = ngroups, nrow = ngroups)
cophw < – rep(0, ngroups)
for (i in 1:ngroups) {
cophw[i] = max(data[which(groups == i), which(groups == i)])
for (j in 1:ngroups){
var[i,j] = max(data[which(groups == i), which(groups == j)])/ cophw[i]
}
}
# Convert distance to similarity
return(1/var);
}
# Script for the symmetric version. This uses the average within group similarity; therefore it can give values greater than 1.
coph3 = function (data, groups, ngroups)
{
var < – matrix(0, ncol = ngroups, nrow = ngroups)
cophw < – rep(0, ngroups)
for (i in 1:ngroups) {
cophw[i] = max(data[which(groups == i), which(groups == i)])
}
for (i in 1:ngroups) {
for (j in 1:ngroups) {
var[i,j] = max(c[which(groups == i), which(groups == j)])/ mean(cophw)
}
}
# Convert distance to similarity
return(1/var);
}
```

B. R Code for PCA Scores Plot PC1 v PC2 for Raw Replicates Tween20 Data

```
Tween1 = read.csv(“Tween20data.csv”, header = FALSE)
View(Tween1)
Tween_info = read.csv(“Tween20info.csv”, header = TRUE)
View(Tween_info)
pcaTween1 = prcomp(Tween1, scale = FALSE)
summary(pcaTween1)
plot(pcaTween1$x[,1], pcaTween1$x[,2], xlab = “PC1”, ylab = “PC2”, pch = 19, col = Tween_info$location)
legend(“bottomleft”, legend = unique(Tween_info$location), pch = 19, col = unique(Tween_info$location))
```

C. R Code for Figure 9 - PCA Scores Plot for Normalised Replicates Tween20 Data

```
Tween2 = read.csv(“Tween20data.csv”, header = FALSE)
View(Tween2)
pcaTween2 = prcomp(Tween2, scale = FALSE)
plot(pcaTween2$x[,1], pcaTween2$x[,2], xlab = “PC1”, ylab = “PC2”, pch = 19, col = Tween_info$location)
legend(“bottomright”, legend = unique(Tween_info$location), pch = 19, col = unique(Tween_info$location))
summary(pcaTween2)
```

D. MATLAB Code for Distribution Fitting Section

```
databrij = ‘BrijCS20mzs.csv’;
BrijCS20mzs1 = readtable(databrij);
datacs20 = ‘BrijCS20data.csv’;
BrijCS20data1 = readtable(datacs20);

tiledlayout(2,3);
nexttile
title(‘Atlas Point Sample’);
plot(BrijCS20mzs1{:,:}, BrijCS20data1{1,:});
hold on;
FWHM1 = 510;
sigma1 = FWHM1/(2*sqrt(2*log(2)));
mu1 = 1214;
x_lim = 0:2800;
fit_Gaussian = normpdf(x_lim, mu1, sigma1);
max_height = 116.74;
lambda = max_height/(max(fit_Gaussian));
z = fit_Gaussian.*lambda;
plot(x_lim,z);
hold on;
FWHM2 = 530;
sigma2 = FWHM2/(2*sqrt(2*log(2)));
mu2 = 1190;
fit_Gaussian2 = normpdf(x_lim, mu2, sigma2);
max_height2 = 230;
lambda2 = max_height2/(max(fit_Gaussian2));
z2 = fit_Gaussian2.*lambda2;
plot(x_lim,z2);

nexttile
title(‘Mevisa Sample’);
plot(BrijCS20mzs1{:,:}, BrijCS20data1{13,:});
hold on;
FWHM1 = 540;
sigma1 = FWHM1/(2*sqrt(2*log(2)));
mu1 = 1240;
x_lim = 0:2800;
fit_Gaussian = normpdf(x_lim, mu1, sigma1);
max_height = 119.48;
```

```
lambda = max_height/(max(fit_Gaussian));
z = fit_Gaussian.*lambda;
plot(x_lim,z);
hold on;
FWHM2 = 545; sigma2 = FWHM2/(2*sqrt(2*log(2)));
mu2 = 1220.02;
fit_Gaussian2 = normpdf(x_lim, mu2, sigma2);
max_height2 = 206.8;
lambda2 = max_height2/(max(fit_Gaussian2));
z2 = fit_Gaussian2.*lambda2;
plot(x_lim,z2);

nexttile
title(‘Rawcliffe Sample’);
plot(BrijCS20mzs1{:,:}, BrijCS20data1{23,:});
hold on;
FWHM1 = 510;
sigma1 = FWHM1/(2*sqrt(2*log(2)));
mu1 = 1220;
x_lim = 0:2800;
fit_Gaussian = normpdf(x_lim, mu1, sigma1);
max_height = 242; lambda = max_height/(max(fit_Gaussian));
z=fit_Gaussian.*lambda;
plot(x_lim,z);
hold on;
FWHM2 = 490;
sigma2 = FWHM2/(2*sqrt(2*log(2)));
mu2 = 1199;
fit_Gaussian2 = normpdf(x_lim, mu2, sigma2);
max_height2 = 97.32;
lambda2 = max_height2/(max(fit_Gaussian2));
z2 = fit_Gaussian2.*lambda2;
plot(x_lim,z2);

nexttile
title(‘Singapore Sample’);
plot(BrijCS20mzs1{:,:}, BrijCS20data1{33,:});
hold on;
FWHM1 = 495.75;
sigma1 = FWHM1/(2*sqrt(2*log(2)));
mu1 = 1210;
x_lim = 0:2800;
fit_Gaussian = normpdf(x_lim, mu1, sigma1);
max_height = 234.07;
lambda = max_height/(max(fit_Gaussian)); z = fit_Gaussian.*lambda;
plot(x_lim,z);
hold on;
FWHM2 = 495.02;
sigma2 = FWHM2/(2*sqrt(2*log(2)));
mu2 = 1190;
fit_Gaussian2 = normpdf(x_lim, mu2, sigma2);
max_height2 = 121.42;
```

```
lambda2 = max_height2/(max(fit_Gaussian2));
z2 = fit_Gaussian2.*lambda2;
plot(x_lim,z2);
```