South African cranial variation: A combined metric-macromorphoscopic method for ancestry estimation

Submitted by:

Leandi Liebenberg

A thesis submitted to the Department of Anatomy, School of Medicine, Faculty of Health Sciences, University of Pretoria, in fulfilment of the requirements for the degree

PhD in Anatomy

2023

Supervisor: Prof E.N L'Abbé

Co-supervisor: Prof K.E Stull

DECLARATION

I, Leandi Liebenberg, declare that this dissertation is my own work. It is being submitted for the degree of PhD in Anatomy at the University of Pretoria. It has not been submitted before for any other degree or examination at this or any other institution.

Sign: 

20 November 2023

Dissertation supervisor

Prof Ericka N. L'Abbé

Professor: Biological Anthropology

Forensic Anthropology Research Centre

Department of Anatomy

Faculty of Health Sciences

University of Pretoria

Dissertation co-supervisor

Prof Kyra E. Stull

Assistant Professor

Department of Anthropology

University of Nevada, Reno

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

Institution: The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

- FWA 00002567, Approved dd 18 March 2022 and Expires 18 March 2027.
- IORG #: IORG0001762 OMB No. 0990-0278 Approved for use through August 31, 2023.

Faculty of Health Sciences **Research Ethics Committee**

16 February 2023

Approval Certificate
Annual Renewal

Dear Miss L Liebenberg,

Ethics Reference No.: 770/2018 – Line 4
Title: South African cranial variation: A combined metric-macromorphoscopic method for ancestry estimation.

The **Annual Renewal** as supported by documents received between 2023-01-18 and 2023-02-15 for your research, was approved by the Faculty of Health Sciences Research Ethics Committee on 2023-02-15 as resolved by its quorate meeting.

Please note the following about your ethics approval:

- Renewal of ethics approval is valid for 1 year, subsequent annual renewal will become due on 2024-02-16.
- Please remember to use your protocol number (770/2018) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, monitor the conduct of your research, or suspend or withdraw ethics approval.

Ethics approval is subject to the following:

- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

On behalf of the FHS REC, Dr R Sommers
MBChB, MMed (Int), MPharmMed, PhD
Deputy Chairperson *of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria*

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health)

Research Ethics Committee
Room 4-60, Level 4, Tswelopele Building
University of Pretoria, Private Bag x323
Gezina 0031, South Africa
Tel +27 (0)12 356 3084
Email: deepeka.behari@up.ac.za
www.up.ac.za

Fakulteit Gesondheidswetenskappe
Lefapha la Disaense tša Maphelo

ABSTRACT

Ancestry is a fundamental parameter of the biological profile. To date South African forensic anthropologists are only able to successfully apply a metric approach to estimate ancestry from skeletal remains. While a non-metric, or macromorphoscopic (MMS) approach exists, limited research has been conducted to explore its use in a South African population. The method has not been sufficiently tested and validated which is required for anthropological methodology to be compliant with standards of best practice. This study aimed to explore the MMS traits and its covariation with cranial measurements to develop improved methodology for the estimation of ancestry from skeletal remains in South Africa. A suite of 17 MMS traits and 25 standard linear measurements were collected from 660 crania of black, white and coloured South Africans.

Inter- and intra-observer agreement was closely scrutinized as visual methods have been shown to be prone to error. The intra-observer agreement ranged from moderate to perfect, with three traits (inferior nasal margin, nasal bone shape, and nasal overgrowth) yielding slightly lower repeatability. Inter-observer agreement was assessed among five individuals with varying levels of general experience and familiarity with the traits. Overall, the observers demonstrated poor to substantial agreement. A group discussion on the scoring procedure, followed by subsequent rescoring of the crania showed a slight increase in overall agreement, with kappa values ranging between moderate and substantial. While general experience does not appear to translate to proficiency with the method, familiarity with the traits and scoring procedure contributes to consistent scores. Thus, method-specific training is essential prior to employing the MMS traits in practice. Technical error of measurement was used to assess the repeatability of the measurements, where the intra-observer error was noted to be lower than the inter-observer error. The greatest disparity was observed with the inter-orbital breadth and mastoid height for both the inter- and intra-observer assessments.

The MMS trait frequency distributions revealed substantial group variation and overlap. Ultimately, not a single trait can be considered characteristic of any one population group. Kruskal-Wallis and Dunn's tests demonstrated significant population differences for 13 of the 17 traits. Black and coloured South Africans, and coloured and white South Africans shared similarities for many of the traits, but black and white South Africans did not present with significant overlap for any trait. ANOVA and Tukey's honestly significant difference (HSD) test revealed that all measurements were significantly different for ancestry, except the foramen

magnum length. Substantial variation and overlap were observed for the measurements among all three groups.

Random Forest Modelling (RFM) was used to develop classification models to assess the reliability and accuracy of the variables in identifying ancestry. Models were created for the traits and measurements separately to gauge the discriminatory power of each dataset. A combined model including all data was also created to test if mixed data can better capture cranial variation than individual methods. The MMS model outperformed the metric model, with classification accuracies of 79% and 72%, respectively. Ultimately, the best results were obtained with the mixed model, which yielded an accuracy of 81%. The results indicate that the combination of size and shape data (as quantified with the mixed model) can effectively distinguish between black, white and coloured South Africans despite significant group overlap. Thus, this study has shown the MMS traits to be a valid and tested method, and the population-specific data from this study can be used to add MMS analyses to forensic casework and skeletal analyses in South Africa.

Key words: forensic anthropology; repeatability; observer experience; classification; random forest models; variable importance

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The primary role of a forensic anthropologist in medico-legal investigations is to establish a biological profile from unknown skeletal remains to provide sufficient information for a presumptive identification. The parameters of the biological profile consist of estimations of age-at-death, stature, sex and ancestry, and can only be established with knowledge of skeletal variation within and between populations. Populations are groups with diverse histories influenced by numerous factors, all of which contribute to the patterned distribution of human variation (Ousley et al., 2009; Spradley and Jantz, 2021). The quantification of skeletal variation among populations forms the basis of ancestry estimation, where the estimation of ancestry is considered possible as skeletal variation has been correlated to socially constructed populations around the world (Jantz and Ousley, 2005). However, it is important to acknowledge that the relationship between skeletal morphology and social race is complicated (Dunn et al., 2020). This inherent complexity should be considered in all aspects of research, including terminology, method design, and drawing conclusions when attempting to quantify population variation from the skeleton (Edgar and Pilloud, 2021). Forensic anthropologists have been more cognisant of this fact and aim to enact transformation in how we describe and explore population variation in the discipline.

The cranium is often considered the most accurate skeletal element for the evaluation of ancestry, with craniometry elected as the preferred approach. Numerous studies have assessed craniometric variation among South Africans (İşcan and Steyn, 1999; Franklin et al., 2010; L'Abbé et al., 2013; Stull et al., 2014a; Maass and Friedling, 2019). The use of standard craniometric variables have been found to produce satisfactory results when estimating ancestry with correct classifications up to 73%. However, standard linear measurements mainly quantify size and are frequently unable to effectively capture the shape variation observable in the craniofacial complex. The use of alternative metric methods, such as geometric morphometrics, has recently gained greater popularity amid anthropological research (Spradley and Stull, 2018). Geometric morphometrics entails recording landmark coordinates of complex objects in a three-dimensional space which then produces statistical and graphical outputs primarily using shape information. Shape differences among specimens can be observed as displacement of individual landmarks within the total configuration of the object being assessed (von Cramon-Taubadel et al., 2007). Researchers have noted coordinate-based analyses achieve greater classification accuracies than standard linear metrics, with

approximately 89% correct classifications among three modern South African groups (Slice, 2007; Stull et al., 2014a). While more accurate, the metric approach, and specifically a morphometric approach may not always be feasible.

The application of non-metric visual assessment appears to be an ideal solution, as the method does not require any equipment, is not time consuming, and can assess cranial size as well as shape. However, the use of non-metrics is associated with numerous methodological issues and is known for perpetuating typological thinking in the assessment and understanding of human variation (Hefner, 2009; Plemons and Hefner, 2016). Acceptance of the *Daubert* criteria (*Daubert* v. *Merrel Dow Pharmaceuticals, Inc.*, 1993) as guidelines for best scientific practice initiated a paradigm shift in forensic research. Closer scrutiny has been placed on both traditional and novel methods. The testing and validation of methodology became recognised as an essential step in ensuring that methods report realistic results, express external validity, and demonstrate forensic significance (Christensen and Crowder, 2009). As such, emerging research around the world has attempted to challenge and improve the non-metric approach, now referred to as the macromorphoscopic (MMS) method, inclusive of adding definitions and comparative drawings, employing robust statistical tests, and gauging the accuracy of the method in different populations (Hefner, 2009; Hefner and Ousley 2014; Plemons and Hefner, 2016; Hefner and Linde, 2018). Greater emphasis has also been placed on exploring observer agreement and trait score variation when employing the traits (Klales and Kenyhercz, 2015; Kamnikar et al., 2018).

Along with emerging research modifying and improving the MMS method, researchers have also explored a more holistic approach that combines both metric and macromorphoscopic methods (Hefner et al., 2014; Maier, 2018). A combined metric-macromorphoscopic approach weighs different aspects of cranial size, shape and morphology, and as such is able to capture more between-group variation, ultimately enhancing classification power (Hefner et al., 2014). Additionally, the use of robust machine learning statistical methods (i.e., random forest modelling, artificial neural networks, etc.) has shown to yield accuracies as high as 90%, which rivals coordinate-based techniques (Hefner et al., 2014; Navega et al., 2015).

To date the MMS method has yet to undergo the same level of application and rigorous scientific testing in South Africa. While the frequency of some of the traits have been assessed, its application in classification models for the purpose of forensic analyses has been very limited (L'Abbé et al., 2011; Dinkele 2018). With a lack of population-specific standards,

South African practitioners may rely on North American standards, which is not recommended as differences have been shown to exist between North Americans and South Africans (L'Abbé et al., 2011, 2013; McDowell 2012, 2015; Krüger 2015; Caple and Stephan, 2017). This requires for more work to be done to ensure the method meets international standards for best scientific practice; more specifically, population-specific standards with known error-rates should be created.

The purpose of this study was to explore cranial variation among black, white and coloured South Africans to improve the methodology employed to estimate ancestry. The objectives included assessing observer agreement in collecting MMS and craniometric data, comparing trait score frequencies and measurement means to identify significant group differences, and creating a series of classification models to test the accuracy with which ancestry can be estimated when using the cranium.

CHAPTER 2: LITERATURE REVIEW

Forensic anthropology as a subdiscipline first gained notoriety among scientists in the United States during the 1930s. However, it was not until the 1960s and 70s that forensic anthropology was formally defined, heralding exceptional growth within the field (Dirkmaat and Cabo, 2012). Modern forensic anthropology, a sister-discipline to biological anthropology, is a well-established scientific field with copious annual publications, numerous students globally, and an immense diversification in the studies conducted. The last two decades has witnessed a shift in research from traditional skeletal biology towards adding taphonomy, archaeology and trauma analysis to the scope of objectives for forensic anthropologists (Dirkmaat et al., 2008). Despite new innovations and research foci, the primary interest of a forensic anthropologist continues to be the creation of a biological profile for unidentified individuals, consisting of age-at-death, stature, sex and ancestry. While modern anthropologists attempt ancestry estimates for social reasons, namely the identification of unknown remains, the concept of race has had a much darker and sordid history in the discipline of anthropology.

2.1 Race in biological anthropology

The analysis of human differences and the concept of race has endured since the emergence of the field, and physical anthropology was essentially considered to be synonymous with racial studies (Caspari, 2003). The need to classify humans into races stems from the empirical principals of taxonomy introduced by Linnaeus in the 18th century (Dubow, 1996). The theoretical foundation of racial science initiated debates surrounding the origin of races on the evolutionary scale, with two major theories dividing the field, namely monogenism and polygenism (Littlefield et al., 1982). The monogenic theory contends that all humans originate from a single evolutionary event involving a common ancestor; this view underscores the rapid and extensive migration of different populations (Gill, 1990). Conversely, the polygenic school of thought maintains that human races are separate biological species that developed from disparate evolutionary lineages (i.e., different ancestors) that evolved in a parallel fashion at different paces (Coon et al., 1950). As such, each racial group were believed to be pure, homogenous and discrete “types”, with any similarities among groups assigned to “admixture” or adaptation to similar environmental stimuli (Caspari, 2009). The polygenic idea of race developed alongside the context of slavery and European imperialism, thereby reinforcing the inherent issue of racial inequality (Littlefield et al., 1982). This typological approach was used

to support the notion that individuals of European descent were evolutionarily superior, leading to scientific racism and discrimination (Dubow, 1995).

The stance of anthropologists in southern Africa was not much different from their European counterparts. With the founding of physical anthropology in southern Africa in the 1870s the field experienced advances in two major directions: early hominin evolution, and assessing population differences, particularly among indigenous groups (Tobias, 1985). While formalised racial science was not quite as intensely represented in South Africa as it was among international anthropologists, especially in Europe, racial theory was present to a substantially greater degree than generally acknowledged (Dubow, 1995).

Changes in the socio-political milieu helped to shift the outlook, with many practitioners in anthropology beginning to question the concept of race (Littlefield et al., 1982). During the 1960s more published research provided novel information on genes and inheritance, allowing the notion of polygenism to be questioned. Furthermore, researchers became more vociferous on how human variation is perceived; the search for pure racial types simply could not be reconciled with the fact that only "hybrids" existed in practice (Dubow, 1996). Great debates surrounding the future of skeletal variation studies followed the denial of the existence of biological races. For instance, statements by Brace (1964) alleging the non-existence of populations as a unit of variation as they intergrade (overlap) with one another effected dissatisfaction among many anthropologists. As it were, concepts embedded in scientific discipline (and in the case of race, also entrenched in public opinion) requires critique, being challenged with data, and ultimately to be replaced with more useful concepts to fade away. It would require the remainder of the 20th century to do away with 19th century thinking. The term race became discontinued as it was recognised to be scientifically impractical and carried harmful social connotations and implications (Lieberman et al., 2003). With the distribution of publications asking questions like "If races do not exist, why are forensic anthropologists so good at identifying them?" (Sauer, 1992), a new modernised outlook overtook the issue of successfully capturing human variation. Anthropologists established that clear differences existed among groups, but the differences were certainly not discrete (acknowledging substantial overlap), and robust methodology was required to quantify and interpret the variation in a meaningful way (Dunn et al., 2020).

Numerous anthropologists set out to evaluate the apportionment of skeletal variation across the globe; results indicated correlations between skeletal biology and population groupings

(Relethford, 2002). Granted, population boundaries are fluid and cannot be explained with arbitrary physical traits or behavioural characters. However, by employing more sophisticated methodology that undergoes rigorous testing anthropologists are better able to interpret patterns of human variation. The inherent variation, both metric and non-metric, has also been noted to emerge early in the ontogenetic process of the human skeleton (Viðarsdóttir et al., 2002; Weinberg et al., 2005; Wood, 2015). Further assessment revealed that enough skeletal variation exists to classify populations at much greater accuracy rates than would be observed with random allocation; thereby demonstrating the classification of humans is not a futile exercise (Ousley et al., 2009). The term ancestry (often conflated with social race) became used in lieu of biological race to abolish any ambiguity between the concepts. Ancestry describes skeletal variation observed on the population-level which has been correlated to different populations across the world, rather than the use of arbitrary morphological traits and behavioural characters as is embodied in the traditional concept of biological race. This observed variation has been described as the result of numerous extrinsic factors, inclusive of culture, language, and geography, and persists because of positive assortative mating and socio-political boundaries (Edgar and Hunley, 2009; Ousley et al., 2009). Importantly, forensic anthropologists provide ancestry estimates, not as a vindication of the concept of biological race, but to allow the presumptive identification of unknown skeletal remains (Stull et al., 2021). The estimation of ancestry has been noted to affect other parameters of the biological profile. Numerous studies have observed population-specific differences in levels of sexual dimorphism and the aging process; thus, prior knowledge of ancestry is required to obtain accurate sex and age estimates (e.g., Oettlé and Steyn, 2000; Spradley and Jantz, 2001; Krüger et al., 2015).

However, the debate on human variation and the estimation of ancestry has recently been reignited with new fervour. Following events pertaining to racially fuelled police brutality in the United States, some anthropologists have come to question how effective ancestry estimates are and whether it still serves a purpose in modern forensic anthropology with some arguing that it does more harm than good (Bethard and DiGangi, 2020; DiGangi and Bethard, 2021). DiGangi and Bethard (2021) even called for the complete abolishment of the estimation of ancestry. Ensuing counterarguments indicated that there were greater discussions to be had on the subject yet (Stull et al., 2021). With that, the year 2021 delivered many publications voicing concerns surrounding the concept of ancestry as it stands and how anthropologists can go about reconciling these concerns with the quantification of skeletal variation in a meaningful,

scientifically valid way (e.g., Edgar and Pilloud, 2021; Michael et al., 2021; Ross and Pilloud, 2021; Spradley and Jantz 2021; Tallman et al., 2021). A number of these authors posit that despite the term ancestry replacing the term race, few changes have occurred in how the research is approached (Ross and Pilloud, 2021; Tallman et al., 2021). Tallman and colleagues (2021) present a thoughtful discussion on the use of ambiguous terminology over the years, and state: *"We define 'ancestry' as biogeographically patterned, clinal, genetic variation that is often continentally derived and defined"* (Tallman et al., 2021:74). In other words, ancestry is largely used to indicate major geographical groupings such as African, Asian, and European. Spradley and Jantz (2021) and Ross and Pilloud (2021) share this sentiment, arguing that the term population affinity better describes the parameter that is being explored with current methods aimed at estimating ancestry. However, some practitioners have already taken ancestry to imply variation on the population level, or what is now being referred to as population affinity. For instance, referring to the Fordisc help file (version 1.53) (Jantz and Ousley, n.d.) which affirms: "*What Fordisc estimates may be termed 'ancestry' in the sense that it identifies population differences resulting from the different origin of each reference population's ancestors"* as well as *"…, these differences reflect the different origins and separate histories of each group which can be highly correlated with many social, geographic, temporal, historical or linguistic groupings of populations".* Admittedly, many South African-based anthropologists have acknowledged that clinal variation is not an adequate descriptor of skeletal variation employed for estimates to facilitate personal identification and have been exploring population-specific variation under the term ancestry (e.g., L'Abbé et al., 2011; L'Abbé et al., 2013; Stull et al., 2014a; Liebenberg et al., 2015; McDowell et al., 2015; Maass and Friedling, 2019). Clearly unease about ambiguities in even the most fundamental terminology (i.e., what to name the concept) is a valid concern.

The inconsistencies and improper methodology used to assess ancestry and population variation within the discipline of forensic anthropology is hereby acknowledged. With the conversation on transformation still ongoing, the term ancestry will be used within and throughout this dissertation. Ancestry in this study refers to the skeletally quantifiable differences that exist within and between populations and the grouping of populations across the globe. The term "population" will be used to refer to a group of individuals sharing a geographical area; in other words, the South African population refers to all individuals that classify as South African. The term "population group" will be used to refer to specific subgroups within the South African population based on the social labelling system currently

employed in the country (i.e., the social race); this includes black, white and coloured South Africans.

2.2 The Estimation of Ancestry

Skeletal variation is the result of several complex, inter-related functional matrices that are subject to individual ontogenetic trajectories (Wood, 2015). These differential trajectories lead to the size and shape differences in the skeleton that forms the basis of sexual dimorphism and population variation; the quantification of skeletal variation is the foundation of the biological profile. Two broad approaches exist to facilitate the evaluation of skeletal variation, namely metric and morphological.

The metric approach involves the measurement of continuous variables using standard landmark definitions and measuring instruments. Measurements of the skull are generally used for the estimation of ancestry, a method which stems back as far as 1926 with the coefficient of racial likeness proposed by Pearson (Spradley and Stull, 2018). With time more robust statistical tests and technological developments became available. Currently one of the most widely used methods in forensic anthropology is the application of discriminant analysis using cranial measurements typically performed with the software Fordisc (Jantz and Ousley, 2005). Custom databases have been created to allow the classification of modern black, white and coloured South Africans (L'Abbé et al., 2013; Liebenberg et al., 2015; Krüger et al., 2017). Although the metric approach reports satisfactory results (ranging between 73% and 84% correct classification using the cranium), the method has some limitations (L'Abbé et al., 2013; Stull et al., 2014a). For example, substantial heterogeneity and group overlap has been shown among black and coloured South African groups (Stull et al., 2014a). In other words, the models are frequently unable to confidently distinguish between black and coloured crania, ultimately resulting in misclassification. One of the reasons for this lack of discriminatory power may be the complex interaction between skeletal size and shape, and the way anthropologists currently attempt to assess it. While the linear data captured with calipers mainly assesses size differences, coordinate data can provide an overall better archive of both size and shape (Slice, 2007; Spradley and Jantz, 2016; Spradley and Stull, 2018). Coordinate data is traditionally captured using a digitiser (such as a Microscribe) rather than callipers. However, the use of digital scanning modalities to source data, such as computed tomography (CT) scans and three-dimensional (3D) surface scanners, have become common practice in biological and forensic anthropology over the last decade (Garvin and Stock, 2016; Franklin

and Blau, 2020). Digital data allows for the 3D reconstruction of skeletal features such as crania, which can be used to collect skeletal dimensions by placing landmarks in the areas to be quantified. In addition to collecting standard inter-landmark distances or coordinate data through geometric morphometric (GM) techniques, the scans can also be used to measure volumes and surface areas of bones (Christensen et al., 2018). Shape analyses have been conducted with GM techniques for the purpose of estimating biological parameters and have found that both cranial size and shape provide useful information and should be considered when attempting to distinguish between groups (Stull et al. 2014a; Maass and Friedling, 2019). More specifically, the general location of specific landmarks has been noted to capture variation more effectively than simple linear distance and result in improved classification accuracy, thereby demonstrating the role that shape differences play in cranial variation (Slice 2007; Stull et al., 2014a). Despite all its advantages, the use of GM and digital data is not always feasible, as it requires expensive equipment and training in its operation and the myriad of statistical analyses associated with it.

The use of morphology and MMS may be a sensible solution and beneficial in this regard, as it rapidly quantifies size, shape and variants of skeletal features across the cranium without the need for additional equipment and is not computationally expensive (Hefner et al., 2012). A wide variety of morphological traits are noted in the literature, but not all are used for the same purpose. Certain minor skeletal variants, such as extra-sutural bones or foramina variations, are commonly referred to as epigenetic traits (Berry and Berry, 1967; Corruccini, 1974; Ossenberg, 1976). The epigenetic traits are dichotomous in nature (i.e., observed as present or absent) and are mainly used in biological anthropology to assess biological affinity on a global scale and population group history (Parr, 2005; Pink et al., 2016). While these traits can provide some information on group relatedness, it is not particularly useful in a forensic context. Instead, the term macromorphoscopic (MMS) trait was proposed to refer to quasi-continuous (i.e., a range of variation rather than present/absent) non-metric features used to assess the ancestry of a single individual for the purpose of forensic identification (Hefner et al., 2012; Pink et al., 2016). MMS traits, which forms the basis of the current study, involves the evaluation of bone shape, bony feature morphology, presence or absence of a trait, as well as feature prominence (Hefner, 2009).

2.3 The Evolution of Non-metrics and Macromorphoscopics

The morphological approach has its roots in the 20th century, particularly through the work of E.A Hooton (1887 - 1954). Hooton was a typologist with an interest in skeletal biology and was known for citing polygenist theory to explain racial variation from the human skeleton (Caspari, 2009). During his tenure at Harvard, Hooton attempted to identify a suite of traits thought to be discriminatory of race; these traits were later compiled into what is known today as the Harvard list. The Harvard list consisted of around 102 observations divided into categories representing the racial classifications in use at the time (i.e., *negroid*, *caucasoid* and *mongoloid*) (Brues, 1990). While popular in application, the trait-list approach proved problematic as the identification of these morphological variants were noted to rely heavily on the experience of the observer, resulting in major interpretive issues. Furthermore, no scientific basis existed for weighing the traits, with personal preference often dictating the choice of traits used to provide an ancestry estimate (Klepinger, 2006; Christensen et al, 2013). Inevitably, the trait-list approach was commonly applied on a *post-hoc* basis to justify an ancestry estimate based on the opinion of the observer rather than describing actual observed cranial variation (Pink et al., 2016; Kamnikar et al., 2018). Simply put, if the observer decided that the cranium in question belonged to a black individual, they only needed to identify one or two traits consistent with black individuals as established by the trait list to support their answer, even if there was evidence to suggest the contrary. Cranial morphology is multivariate, and the cranium can be broken down into separate units that variably reflect size and shape differences attributable to population history or environmental influence (von Cramon-Taubadel, 2014). Therefore, the ability to identify ancestry is not based on the identification of single, univariate traits, but by viewing the skull in its entirety, as a skull rarely, if ever, portrays traits from only one population (Hefner et al., 2012).

The non-metric methodology was recognised to be fraught with subjectivity issues; an unpublished manuscript by Hooton testing the trait-list approach revealed a mere 18.7% agreement among observers (Hefner et al., 2004). Although Hooton addressed the need for standardisation of the method, the approach remained largely unchanged throughout the years (Brues, 1990; Hefner, 2009). Hooton's theoretical approach to race, and in so doing his Harvard list, was adopted by many of his students in their own research (Hefner et al., 2012; Hefner, 2018). Passed on from mentor to student through oral tradition the morphological trait-list approach remained deeply entrenched in physical anthropology and the evaluation of human variation. One of the most frequently cited manifestations of the non-metric approach

completed by Rhine (published as recent as 1990) is essentially a version of Hooton's Harvard list (Hefner, 2003). The non-metric method for estimating ancestry employed by Rhine (1990) is based on extremely small samples (some as little as nine individuals used to represent an entire population), extreme trait expressions allowing for little group overlap, and no statistical analyses. Furthermore, lack of standardisation in the application of the method has seen this typological, unscientific approach persist into the legacy of modern forensic anthropology for much longer than it should have.

After the introduction of the *Daubert* criteria (*Daubert* v. *Merrel Dow Pharmaceuticals, Inc.*, 1993) the method, and other methods with similarly questionable application, began to receive much needed scrutiny. The *Daubert* criteria for best scientific practice were implemented following a case of United States Federal legal proceedings with a dubious outcome as a result of expert testimony (Grivas and Komar, 2008; Christensen and Crowder, 2009). Following the guidelines set forth by the *Daubert* criteria, expert witness testimony is required to be substantiated with scientifically tested methods that have been shown to be repeatable and precise and produce error rates and probability assessments (Dirkmaat et al., 2008). This motion thereby rendered investigator experience insufficient as a justification for a scientific conclusion. While South African courts do not specifically adhere to the *Daubert* criteria, similar guidelines should be followed to place forensic science in South Africa on par with international standards of best practice (Allan and Louw, 2001; Meintjes-van der Walt, 2003). More recent publication of reports by the National Academy of Sciences (NAS, 2009) and the President's Council of Advisors on Science (PCAST, 2016) in the United States has reiterated the need for specialised validation of methodology employed in forensic practice.

Despite the resounding affirmations of the *Daubert* criteria in the early 1990's, the traditional trait-list approach for ancestry estimation persisted several years longer before researchers began to transform it. In 2003, Hefner published a master's thesis addressing the trait-list approach; this was the first in a long line of studies modernising the method to be *Daubert* compliant. The transformation of the method also welcomed a new label, as non-metric gave way to macromorphoscopic. The study introduced a multi-state scale to describe the traits as it provides a more realistic reflection of how the traits are expressed than a binary scale. The ordinal scale was paired with comprehensive descriptions and comparative illustrations to depict the different trait states. This methodology is consistent with other popular morphoscopic techniques used to evaluate sex from the cranium and pubic bone (Walker 2008; Klales et al., 2012). The use of comparative line drawings and descriptive

definitions have assisted to counteract some of the discrepancies that stem from observer error associated with non-metric methods.

Ultimately Hefner (2003, 2009) observed much lower trait frequencies than traditionally assumed to be present, with numerous traits failing to demonstrate significant differences among groups. Further additions by Hefner (2007) and Hefner and Ousley (2014) gauged the application of a variety of classification methods to create models for estimating ancestry. Thus far the method has been expanded to also include other modern groups, such as Hispanics and Asians, in addition to the black and white North Americans included in the original study (Hefner et al., 2015; Plemons et al., 2018; Maier and George, 2020). To allow standardised data collection a software program, Osteoware® (Smithsonian Institution, 2011), incorporated a module for scoring MMS traits. Osteoware® also includes modules for collecting data pertaining to other areas of skeletal biology, such as craniometrics and pathology, amongst others. The module was later adapted to become a freestanding program (*MMS* 1.6.1), which was developed with the specific purpose of collecting MMS data. Similar to the Osteoware module, *MMS* is a graphical user interface (GUI) that provides a user window with the definition and each trait state along with illustrations and descriptive notes to streamline the scoring procedure (Kamnikar et al., 2018). The latest contribution to the above-mentioned research is the creation of the MMS databank (MaMD). Similar to the Forensic Databank of measurements, the MaMD is designed for the collection and analysis of MMS data worldwide to create more appropriate reference samples for research. The MaMD currently contains data from more than 2300 modern individuals and encourages data sharing and collaboration among researchers (Plemons and Hefner, 2016; Hefner 2018). The MaMD analytical tool (version 0.3.15) has also been made available to classify unknown crania into a reference group from the database using artificial neural networks (http://macromorphoscopic.com/).

2.4 Macromorphoscopics in South Africa

Research pertaining to the non-metric approach in South Africa have been scarce. In the past South African anthropologists have mainly relied on the work of De Villiers (1968) supplemented with international standards when conducting morphological analyses (Krogman, 1962). This practice made use of outdated information, was blind to human variation as it lacked population-specific data and was ultimately typological. Following the innovations by Hefner, L'Abbé et al. (2011) conducted a pilot study to assess the use of macromorphoscopic traits on modern black, white and coloured South Africans. The study

aimed to assess the prevalence of thirteen macromorphoscopic traits; nine traits were taken from Hefner's (2009) suite of traits, combined with an additional four traits adapted from Bass (1995) and Hauser and de Stefano (1989). Tests to gauge the observer variation obtained agreement levels ranging from moderate to excellent for most traits. However, six traits proved to be difficult to score consistently; this notably included the four traits not modified by Hefner, thereby emphasising the importance of the illustrations in assigning the correct trait state. Frequency distributions revealed near equal distributions of the traits among all three groups. The lack of group separation was ascribed to the inherent heterogeneity of the South African population (L'Abbé et al., 2011). The study did not create any classification models, thus the accuracy with which the MMS traits could classify South African remains speculative.

With such moderate results the non-metric approach has subsequently been omitted from South African forensic anthropological analyses. Unfortunately, no other research has been published to further explore the MMS variation among South Africans. However, the recent inclusion and successful classification of additional groups into the MaMD (i.e., North American Hispanics - a group previously recognised to be highly heterogeneous) suggests that the method may be more suitable to the South African population following some methodological adjustments and the inclusion of robust statistical models to better interpret the subtle trait compositions among the groups (Hefner et al., 2015).

2.5 Method validation and methodological considerations

Validation studies are essential to the advancement and standardisation of the field of biological anthropology. Studies revising methodology can assist in contributing to a better understanding of limitations and biases associated with methods and is paramount to yield reliable results. Given the presumed subjective nature of non-metric methods, trait score variation and its implications on correctly classifying biological parameters continues to be a concern that warrants further study (Hefner, 2009; Klales et al., 2020). This problem has prompted many authors to dedicate entire papers to quantifying sources of observer error in study topics including pathology, aging, and the estimation of sex and ancestry (e.g., Shirley and Ramirez Montes, 2015; Wilczak et al., 2017; Kamnikar et al., 2018; Klales et al., 2020).

Once scientifically acceptable methodology was in place other researchers also began to explore the MMS traits. Klales and Kenyhercz (2015) conducted a validation study to evaluate the amended method. While their study confirmed the external validity of the MMS approach, some points of concern and areas for improvement were addressed. One of the most prominent

deductions involved the need for training with the traits prior to employing the method in practice. Observer experience has consistently been identified as one of the greatest sources of discrepancies in trait scores in morphoscopic techniques (Wilczak et al., 2017; Klales et al., 2020). Published figures and descriptions have been developed in such a way that theoretically anyone should be able to use morphoscopic methods to assign the score that matches most closely (Klales et al., 2020). However, this is not always the case and validation studies frequently report poorer results than the original publications. One possible explanation for this is that original studies usually involve the developer of the method, whether it is directly through the data collection itself, or indirectly through training (Wilczak et al., 2017; Klales et al., 2020). As such the reported results may underestimate any issues that arise with scoring. Indeed, reproducibility testing is essential, but results from independent researchers give a more realistic reflection of the method as it would be used in practice.

Studies have found that individuals with greater levels of general experience with skeletal material tend to produce more consistent results, regardless of method-specific experience (Wilczak et al., 2017; Kamnikar et al., 2018; Klales et al., 2020). This has been ascribed to more experienced practitioners having been exposed to more human variation which allows them to identify and recognise more subtle skeletal differences more effectively than their less experienced counterparts. However, whether general experience in the field translates to scoring competency may depend on the method itself. When assessing traits of the cranium and pubis for sex estimation (Walker 2008; Klales et al., 2012), Klales and colleagues (2020) report that more generally experienced observers produced the best results. But even though observers with less general experience were more variable in their scores, the results still demonstrated good agreement overall. Thus, knowledge on skeletal variation may contribute more to application of these methods than formal training on the specific scoring systems (Klales et al., 2020). This contrasts somewhat to results obtained by Klales and Kenyhercz (2015) when assessing the MMS traits to estimate ancestry. Although the authors also noted that the two experienced observers produced more consistent scores, the agreement for most of the traits assessed were slight to moderate, indicating the need for method-specific training prior to scoring the traits (Klales and Kenyhercz, 2015). Kamnikar et al. (2018) echoes this statement, as their results exploring long-term intra-observer trends in scoring the same traits revealed that greater experience with the method leads to less extreme trait scores.

A possible reason for the need for prior training for the MMS traits could be the greater number of traits with a more variable, more complex scoring system compared to the sex

estimation scoring methods. For sex, the Walker (2008) method employs five traits on the skull and the Klales et al. (2012) method employs three traits on the pubic bone. Both methods make use of an ordinal scale ranging between one and five for all traits with ranked scores (i.e., 1>2>3>4>5). Conversely, the MMS traits for population affinity started with 11 traits in the original publication (Hefner, 2009), and now includes 17 traits in the latest texts (Hefner and Plemons, 2016; Hefner and Linde, 2018). Furthermore, the codification varies among each of the traits, ranging from ordinal with ranked scores (e.g., nasal aperture width); nominal, where there are separate categories with no particular ranking (e.g., nasal aperture shape); and binary, where the trait is either present or absent (e.g., nasal overgrowth) (Hefner and Linde, 2018). The intricacies of the scoring system may lead to more discrepancies in the trait scores, especially if an individual has no prior training or experience with the traits.

The methodology used in validation studies may also play a role in score variability, rendering results incomparable between studies. Varying samples and the number of trials or observers can all affect the results. For instance, smaller samples may affect the prevalence of certain traits; if a sample only contains crania that lack any post-bregmatic depressions, the results may demonstrate perfect agreement between observers, but does not accurately reflect the ability of the observer to correctly identify and score the trait. The choice of statistics and how it is applied may also affect results. Cohen's kappa is widely used by forensic anthropologists to compare scores. However, there is not always agreement in which variations of kappa should be used; more specifically, whether the traits should be weighted or not. Weighted kappa calculations may be better suited to test the agreement of ordinal ranked traits, as the weighted kappa allows the user to consider the importance of disagreements between scores; in other words, how harshly a disagreement in scores should be penalised (Kamnikar et al., 2018; Tran et al., 2018). Sim and Wright (2005) recommend the use of the weighted kappa for ordinal data and cautions against comparing kappa values across variables with different prevalence or bias, or traits that are measured on different scales. The use of different methods and tests and how it affects the results of anthropological analyses has been touched on (Klales et al., 2021), but is certainly not discussed enough. With the MMS traits, many studies have made use of the traditional unweighted Cohen's kappa (e.g., Hefner, 2009; L'Abbé et al., 2011; Klales and Kenyhercz, 2015), with a few more recent written works arguing for the use of a weighted Cohen's kappa (e.g., Kamnikar et al., 2018; Maier, 2018; Merchant, 2023). However, there is limited comparative data available to gauge the effects of different choices in statistics on the results of observer testing.

The statistical analysis of categorical data on which MMS is based is notoriously complex; however, the availability of software packages allows for simpler application of robust statistical support. Although a variety of classification statistics can be used, non-parametric methods are recommended as ordinal variables may violate the assumptions associated with parametric methods (Pink et al., 2016). The violation of assumptions will likely lead to poor model performance, low classification accuracies, and difficulty interpreting the results. Therefore, methods commonly employed with anthropological analyses (such as discriminant analysis and logistic regression) are not expected to deliver an optimal performance when using ordinal data. More robust machine learning techniques have been proposed as alternative options to analyse the data. Machine learning involves tuning a large number of random cut-off points in a sample to identify the most discriminatory way to group individuals (Hefner and Ousley, 2014). Some examples of machine learning techniques explored in anthropology are artificial neural networks (ANN), support vector machines (SVM), and random forest models (RFM). Hefner and Ousley (2014) compared the discriminatory performance of numerous methods, including both parametric and non-parametric machine learning methods; with a three-group classification ANN (87.8%), SVM (86.4%) and RFM (85.5%) obtained the highest overall accuracies. With such similar accuracies other factors need to be considered when choosing a method for model creation, namely computational requirements and difficulty of interpretation. All three of these machine learning methods have been labelled as computationally expensive and requires extensive user experience to interpret the results (Huang et al., 2004). However, RFM has some advantages over ANN and SVM that makes it more appealing. First, RFM has a variable importance measure that simplifies output interpretation by indicating which variables contribute most to the model (i.e., which traits help to separate the groups) (Fabris et al, 2018). Additionally, RFM is able to analyse a combination of both categorical and continuous data, making it ideal for assessing combined metric-macromorphoscopic models (Zheng et al., 2009). RFM is not a new method to biological anthropology and the positive results from previous studies warrant its use for the current study (Hefner and Ousley, 2014; Hefner et al., 2014; Navega et al., 2015).

CHAPTER 3: MATERIALS AND METHODS

3.1 Sample

The South African population is diverse and consists of four major groups: South African blacks (81.0%), whites (7.7%) and coloureds (8.8%) make up the majority of the population; the remaining 2.6% of the population consists of individuals classified as Asian and Indian (Statistics South Africa, 2022). Each group has a unique history within the country leading to the vast heterogeneity observed within and among the groups. Black South Africans descend from Bantu-speaking groups that migrated throughout sub-Saharan Africa from western-central Africa approximately 3000 to 5000 years ago (Tishkoff and Williams, 2002). Further divisions among the southern Bantu-speakers based on factors associated with kinship, religion and language resulted in the numerous subgroups residing in southern Africa today, inclusive of Nguni, Sotho, Venda and Shangaan-Tsonga (Stull et al., 2016). Colonisation of the Cape during the 17th century introduced European settlers to South Africa, shaping the heritage of white South Africans. The settlers were mainly of Dutch origin, with additional contributions from French Huguenots and Germans that arrived in the 18th century. Late in the 18th century South Africa was also colonised by the English (Liebenberg et al., 2015). Coloured South African refers to a self-identified group unique to South Africa. The group is a result of the complex history of South Africa with genetic contributions from Khoe-San (considered indigenous South Africans), Bantu-speakers, Europeans, as well as Indians and other Asian groups that were brought to South Africa as slaves to maintain the Cape colony. The complex population structure and history of the coloured South Africans manifests as a genetically and skeletally heterogeneous group with substantial variation (Stull et al., 2014a). While the varying origins of each group resulted in a uniquely heterogeneous population with distinct structures, the group differences employed to attempt ancestry estimations persisted as a result of socio-political boundaries. Sociocultural identity in South Africa is based on the categorisations assigned to individuals during the *Apartheid* era, which contributed to widespread endogamy among groups (Krüger et al., 2018).

The sample consisted of 660 crania (n = 220 black, white, coloured South Africans), with equal sex distribution. The crania were sampled from the Pretoria Bone Collection (University of Pretoria) and the Kirsten Collection (Stellenbosch University). The skeletal material housed in the stated collections are derived from cadavers obtained from either donated or unclaimed

bodies received under regulation of the medical schools of the respective institutions. The remains accessioned into the collections are of documented sex, age-at-death, and peer-reported ancestry (L'Abbé et al., 2005; Alblas et al., 2018). Ethical approval (770/2018) to conduct the study was obtained from the Faculty of Health Sciences Research Ethics Committee at the University of Pretoria. Due to the presence of post-mortem damage and/or ante-mortem trauma not all of the traits and measurements could be collected for each cranium. As such the sample size differs for each variable. Table 3.1 provides the mean age distribution for each group. Notably the mean age for the white South Africans is older than the other groups. While age was not specifically investigated in this study, it should be acknowledged as a potential covariable, particularly when age differences are present among the groups in the sample. Reports have shown that the prevalence of tooth loss and edentulism increases markedly in the South African population after the age of 45 years (Kimmie-Dhansay et al., 2021). This was certainly the case in the sample, with many individuals being either partially or completely edentulous, which may have implications on the data collection and results.

Table 3.1 – Age distribution (years) of the skeletal sample.

	Black			**White**			**Coloured**		
	Males (n=110)	Females (n=110)	Pooled (n=220)	Males (n=110)	Females (n=110)	Pooled (n=220)	Males (n=110)	Females (n=110)	Pooled (n=220)
Mean	49	46	48	62	68	65	48	47	47
Range	19 - 98	16 - 75	16 - 98	31 - 85	21 - 97	21 - 97	19 - 85	18 - 100	18 - 100

3.2 Methods

The data collection for the study was comprised of two components, namely (1) MMS and (2) metric, which will be discussed separately.

3.2.1 MMS component

A total of 17 MMS traits were visually assessed on each cranium. Refer to Table 3.2 for a list of the macromorphoscopic traits included in the study, and Appendix I for the trait definitions. Each MMS trait was scored following the latest methodology proposed by Hefner et al. (n.d.) and Plemons and Hefner (2016) as used in the "Macromorphoscopics Software" data collection module (*MMS* version 1.6.1). *MMS* is a free-standing version of the data collection software *Osteoware*® (Smithsonian Institution, 2015) created specifically to facilitate the collation of a global MMS databank (MaMD) (Hefner, 2018). Data entry in *MMS* includes definitions with accompanying line drawings demonstrating the different variable expressions, or states, for each trait (Plemons and Hefner, 2016). Different trait expressions receive scores organised on a scale; i.e., depending on the number of states available, each trait will be given a score based on a binary (two states) or ordinal/nominal (more than two states) scale. Where traits are bilaterally expressed, only the left side was scored. In instances where the left side was not available, the right side was used. Hefner and colleagues (n.d.) recommend the anterior nasal spine not be scored for edentulous individuals. The anterior nasal spine was still scored in this study despite ante-mortem tooth loss. This should be considered when interpreting the results; but ultimately the implications of tooth loss on the size and shape of the anterior nasal spine need to be further explored.

Table 3.2 – Macromorphoscopic traits and abbreviations.

Anterior nasal spine	**ANS**	Nasofrontal suture	**NFS**
Inferior nasal aperture	**INA**	Orbital shape	**OS**
Interorbital breadth	**IOB**	Post-bregmatic depression	**PBD**
Malar tubercle	**MT**	Posterior zygomatic tubercle	**PZT**
Nasal aperture shape	**NAS**	Supranasal suture	**SPS**
Nasal aperture width	**NAW**	Transverse palatine suture	**TPS**
Nasal bone contour	**NBC**	Palate shape	**PS**
Nasal bone shape	**NBS**	Zygomaticomaxillary suture	**ZS**
Nasal overgrowth	**NO**		

3.2.2 Metric component

A total of 25 standard linear cranial measurements were collected from the same crania used in the MMS component. All measurements were taken to the nearest millimetre using a standard manual sliding caliper and spreading caliper following the definitions of Langley et al. (2016). Refer to Table 3.3 for a list of the measurements included in the study, and Appendix II for the measurement definitions. Measurements were only taken on the left side. If the left side was unavailable, the right side was used. Measurements surrounding the palate (MAL and MAB) were not collected in instances of substantial alveolar resorption due to ante-mortem tooth loss.

Table 3.3 – Measurements and abbreviations.

Maximum cranial length	**GOL**	Nasal breadth	**NLB**
Maximum cranial breadth	**XCB**	Orbital breadth	**OBB**
Bizygomatic breadth	**ZYB**	Orbital height	**OBH**
Basion-bregma height	**BBH**	Biorbital breadth	**EKB**
Basion-nasion length	**BNL**	Interorbital breadth	**DKB**
Basion-prosthion length	**BPL**	Frontal chord	**FRC**
Maximum alveolar length	**MAL**	Parietal chord	**PAC**
Maximum alveolar breadth	**MAB**	Occipital chord	**OCC**
Biasterionic breadth	**ASB**	Foramen magnum length	**FOL**
Upper facial height	**NPH**	Foramen magnum breadth	**FOB**
Minimum frontal breadth	**WFB**	Mastoid height	**MDH**
Upper facial breadth	**UFBR**	Biauricular breadth	**AUB**
Nasal height	**NLH**		

3.3 Statistical Analyses

All statistical analyses were completed using the software R version 4.1.0 (R Core Team, 2021). Outliers were detected and removed prior to further analysis. Univariate boxplots were used to screen the MMS data, and a combination of univariate boxplots and bivariate scatterplots were used to screen the metric data. Additionally, the metric data was subjected to additional analyses to determine if the data meets the assumptions associated with parametric tests. A Shapiro-Wilk test was used to assess normality (Appendix III) and a Levene's test was used to assess homoscedasticity of variance (Appendix IV). Non-parametric statistics have previously been recommended to assess MMS traits as the ordinal scale of the character states may violate typical parametric statistical assumptions (Pink et al., 2016). Thus, suitable non-parametric statistics were selected, and no additional *ad hoc* tests were conducted for the MMS data.

3.3.1 Observer agreement

For anthropologists to confidently create standards and employ methods, both the precision and repeatability (intra-observer agreement) as well as the reliability (inter-observer agreement) of the method needs to be tested. Inter-observer agreement evaluates consistency in observations of the same feature between two or more individuals, while intra-observer agreement gauges consistency over multiple attempts by a single observer (Kamnikar et al., 2018). The visual assessment of traits is notoriously subjective and prone to increased levels of observer error (Hefner, 2009). Even with the addition of line drawings and more descriptive definitions, studies report variable – and often poor – observer agreement (L'Abbé et al., 2011; Klales and Kenyhercz, 2015). The use of different statistics, varying samples, population differences, and differences in the traits being assessed means that the results of validation studies are not always directly comparable to one another. For this reason, a detailed analysis of observer agreement and trait score variation was conducted for the MMS component of the study. Craniometry is considered more objective with less influence from external factors, and has been thoroughly assessed (Langley et al., 2018; Smith and Boaks, 2018; Liebenberg and Krüger, 2020). As such, the repeatability of the measurements was not tested as rigorously as the MMS traits.

3.3.1.1 MMS component

The sample consisted of 10 crania selected from the Pretoria Bone Collection. To ensure a wide variety of trait expressions, thus avoiding statistical issues with trait prevalence, the sample included black and white South African males and females. The demographic information was not disclosed to the observers to prevent any potential cognitive bias. The 17 MMS traits were scored on each cranium by five different observers. The observers differed in their levels of experience regarding osteology and forensic anthropology as well as their experience with the traits. Table 3.4 provides a summary of the observers, with information on their level of education and number of years they have been involved with forensic casework, data collection and osteological research at the time the data were collected. Observer A (the principal investigator) has extensive experience in the field, having worked with skeletal material for 10 years, which includes forensic case analysis, data collection and teaching. Additionally, they also have extensive experience with the traits, having received training (from an expert with experience using the traits, but not a method developer), and self-trained with the figures and descriptions for approximately three years prior to data collection. Observer B is the only other participant familiar with some of the traits, having published on the subject. Observers C – E have no experience with the traits and vary in their general experience.

For the scoring, each observer was supplied with the *MMS* software and the *MMS* user guide. As per the recommendation of the MMS user manual, a contour gauge was provided to assist with the scoring of the nasal bone contour and the post-bregmatic depression, and a clear ruler was provided to assist with the scoring of the malar tubercle and the posterior zygomatic tubercle. Only the left side was scored in the case of bilateral traits.

Each of the five observers scored the crania by themselves, without discussing the scores with one another. After all observers had completed the scoring, a group discussion session was held to deliberate on the scoring procedure. During this session the observers went through the descriptions for each trait, and how they each went about assigning a score and resolving scores for any traits they were conflicted about. A series of additional crania (independent of the ones being scored for analysis) were brought to the discussion session to showcase different examples of the traits as well as some variants that may complicate scoring. Each observer then rescored the same crania that were originally scored within a period of four to six weeks after the first round of scores. Once again, the observers scored the crania individually without

discussing their scores. All the same tools (*MMS* software, user manual, contour gauge, and clear ruler) was made available to the observers.

Table 3.4 – Summary of observer experience.

Observer	Highest education/ employment	Trait experience	General experience
A	PhD student, practicing forensic anthropologist	Extensive (scored the traits in > 100 individuals, received training)	10 years' experience working with forensic cases and data collection in SA
B	PhD, practicing forensic anthropologist	Moderate (scored the traits in < 50 individuals)	20 years' experience working with forensic cases and data collection in SA and USA
C	PhD student, practicing forensic anthropologist	Novice (has never scored the traits or used the method)	10 years' experience working with forensic cases and data collection in SA
D	MSc student	Novice (has never scored the traits or used the method)	2 years' experience working with forensic cases and data collection in SA
E	BSc undergraduate student	Novice (has never scored the traits or used the method)	Very limited experience working with skeletal material

The observer agreement was then assessed with Cohen's kappa which was calculated with the *irr* package in R (Gamer et al., 2019). The kappa coefficient measures the agreement between observers in assigning categorical variables adjusted by the standard measure of reliability that could be expected due to chance (Walrath et al., 2004; Ferrante and Cameriere, 2009). Calculated kappa values can range from -1 to 1, where values closer to 1 indicate greater agreement. On the other hand, a negative value indicates agreement due to chance (Walrath et al., 2004). There is currently no universally accepted cut-off point for satisfactory observer agreement. However, to be consistent with nomenclature when describing the strength of

agreement associated with kappa statistics, the parameters proposed by Landis and Koch (1977) were used. The parameters are outlined as follows:

< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

Different weights can be assigned to categorical variables depending on the data structure of the trait (i.e., binary, nominal or ranked ordinal) and how harshly disagreement in a score should be penalised (Sim and Wright, 2005; Tran et al., 2018). While an unweighted kappa is suitable for binary and nominal structured traits (where any score disagreement is equally penalised), a weighted kappa should be considered for ordinal traits that have a specific rank or order to the scores (Sim and Wright, 2005). To better explore the implications of different modifications to the statistical test, a series of analyses were run using different weights for the traits for the intra-observer agreement. This included the traditional unweighted Cohen's kappa for all traits; linear-weighted Cohen's kappa for all traits; and quadratic-weighted Cohen's kappa for all traits. Lastly, a mix of unweighted (for binary and nominal traits) and quadratic-weighted (for ordinal ranked traits) Cohen's kappa was applied to the appropriate traits. For the inter-observer agreement, a mixed Cohen's kappa was selected to compare the scores for each additional observer with the primary observer to explore individual trends. A mean kappa value was then calculated for each trait to see the overall repeatability of the traits when considering all observers simultaneously. A Holm's adjustment was applied to avoid familywise error with multiple comparisons. The inter-observer agreement was calculated both before and after the group discussion to see if more familiarity with the traits and the scoring procedure influenced the agreement.

3.3.1.2 Metric component

The same ten crania were used to test the repeatability of the 25 measurements collected by two observers, namely the principal investigator (observer A) and one additional observer (observer C). The measurements were only repeated once, as both observers had extensive experience with osteometry, and further discussion was not required. The inter- and intra-

observer agreement was assessed using absolute and relative technical error of measurement (TEM and %TEM, respectively). Absolute TEM provides an accuracy index that expresses error margins through the standard deviations of repeated measurements, where a higher TEM value indicates greater variation or error (Perini et al., 2005). The absolute TEM can be converted to %TEM to take the overall size of the measurement into account, as error with a relatively small measurement will have greater implications than the same error with a much larger measurement. Bland-Altman plots were also created to visually demonstrate the measurement agreement and overall variability between observers.

3.3.2 Exploratory analyses

A series of exploratory tests were conducted to test for group differences among black, white and coloured South Africans and to assess the relationship between the MMS and metric variables. Though exploratory analyses were used to evaluate the effects of sex, the objectives of the current study were not to assess sex differences among South Africans. Thus, the sexes were pooled for further analyses unless indicated otherwise.

3.3.2.1 MMS component

The MMS scores were used to create frequency distributions to assess the occurrence of each trait per group. Kruskal-Wallis tests were used to identify if any traits demonstrated significant differences among the populations. Kruskal-Wallis is a non-parametric test used to compare three or more groups which operates under the assumptions of independence of scores but is not bound by assumptions of normality or homogeneity of variance (Lee, 2022). Additionally, a *post-hoc* Dunn's multiple comparisons test (with a Holm's adjustment) was used to further explore differences in the trait frequencies among the populations. The Holm's adjustment counteracts the effects of multiple comparisons and prevents increased probability of Type I errors occurring (Ali and Bhaskar, 2016). More specifically, where Kruskal-Wallis indicates the presence of significant differences, the Dunn's test indicates which groups in a multiple comparison differ from one another to better interpret group overlap.

3.3.2.2 Metric component

An analysis of variance (ANOVA) was performed to compare the group means per measurement. ANOVA is a parametric test that operates under the assumptions of normality, homoscedasticity, and independence of variables (Scariano and Davenport, 1987). The effects

of ancestry, sex and the interaction between ancestry and sex was assessed to identify any potential group differences. A *post-hoc* Tukey's honestly significant difference (HSD) test (with a Holm's adjustment) was performed in conjunction with the ANOVA to further identify which groups demonstrated significant differences from one another. Following the results of the *ad-hoc* Shapiro-Wilk and Levene's tests, a Kruskal-Wallis test (with a *post-hoc* Dunn's test) was used to assess any variables that violate the assumptions of the ANOVA.

3.3.2.3 Correlations

A series of correlation analyses were conducted to assess the relationships among the different MMS traits, as well as the metric variables. More specifically, polychoric correlations were used to assess the relationship among the MMS traits, Pearson correlations were used to assess the relationship among the measurements, and polyserial correlations were used to assess any possible relationships between the MMS traits and the measurements. Correlation coefficients can range between -1 and 1, indicating either a positive or negative relationship. Values closer to ±1 indicate the strength of the relationship, where greater coefficients suggest higher degrees of covariance between variables. The sign of the correlation (+ or -) indicates the nature of the relationship. Positive correlations indicate, for example, that if the value of one variable increases that the value of another variables another increases along with it, while negatively correlated variables have an inverse relationship (Curtis et al., 2016). While there is no universal agreed upon interpretation of correlation coefficients, the following descriptions proposed by Chan (2003) was employed in this study:

< 0.30	Poor
0.30 – 0.59	Fair
0.60 – 0.80	Moderate
> 0.80	Very strong

In addition to assessing associations among variables, correlation coefficients are used to test for multicollinearity. Multicollinearity exists when two independent variables are highly correlated, and can inflate standard errors, bias inference statistics, and lead to unstable parameter estimates, ultimately affecting the interpretation of results (Dormann et al., 2012). Caution must be carried out when interpreting results when multicollinearity is present, which is suggested to occur with correlations greater than 0.9 (Tabachnick and Fidell, 2007). The *polycor* package in R was used to conduct the correlations (Fox, 2022).

3.3.3 Classification models

Random forest models (RFM) were employed to classify the population affinity of the crania. RFM is a non-parametric machine learning method that was introduced as an improvement upon decision trees (Breiman, 2001; Klales and Kenyhercz, 2015). Decision trees are a type of classification model that uses sequential splitting values (such as MMS traits) to predict the probability of an unknown belonging to a certain class (i.e., ancestry) to separate a dataset into groups (Hastie et al., 2009). Within each data split, known as "nodes" in the tree, the variable that is most strongly associated with the response variable (a specific group) is selected for the next split until a stopping condition is met. In the case of the current study, the stopping condition is an overall ancestry estimate based on the ensemble of multivariate trees. The overall ancestry estimate is reached by combining the most likely response from all the nodes, or in the case of RFM, all of the trees in the ensemble. This is achieved by means of voting in classification; simply put, the population group that receives the most "votes" from the trees is returned as the overall prediction (Breiman, 2001). A total of 2500 classification trees were used for each model with four variables at each split. Furthermore, RFM ranks the importance of each variable included in the classification ensemble, giving an indication of which variables are most discriminatory in the model and which variables are being "noisy" and do not contribute to the classification (Hefner and Ousley, 2014). With variable importance, the higher the value, the more a variable contributes to the classification. Finally, out-of-bag observations can be used to gauge the external prediction accuracy of the tree (comparable to leave-one-out cross-validation used with discriminant analysis). The original training data is randomly sampled with replacement for each tree, which generates a smaller subset of data for each tree; essentially this is the training data. The observations excluded from the training data, or the out-of-bag observations, are a random subset of data that is essentially an internal test sample. The tree will then be used to classify the test sample to obtain a more realistic classification accuracy (Strobl et al., 2009).

In the case of missing data, the mode was calculated for each trait per each sex and population group separately. The mode was used as an imputation value specifically because it appears the most in a set of values which in this case, is a population and sex group, most individuals are likely to depict that value. Data imputation was only performed when variables had less than 10% of the observations missing. For variables where more than 10% of the observations would have to be replaced, the variable was omitted from the model. After the missing data were imputed, the sample was divided so that 75% was used as the training set to

create the model, and the remaining 25% was the holdout set to test the accuracy of the model on an independent set of crania.

Both univariate and multivariate analyses were conducted to evaluate the performance of the traits when tested both individually and in combination. Overall, three multivariate models were created, namely (1) an MMS model; (2) a metric model; and (3) a combined MMS-metric model. The classification accuracy (for both the training and testing samples), Kappa values, and variable importance were recorded for each model. Both the classification and Kappa values are measures of model accuracy. The classification accuracy presents the percentage of correctly classified individuals out of all the individuals; whereas, the Kappa value presents the percentage of correctly classified individuals while taking random chance into account. The *randomForest* package was used to generate the RFM classifications (Liaw and Wiener, 2002).

CHAPTER 4: RESULTS - MORPHOSCOPIC VARIATION

4.1 Observer agreement

The intra-observer agreement was assessed using Cohen's kappa with varying weights assigned to the traits (Table 4.1). The mean kappa value varies depending on which weights are applied, with the unweighted kappa producing the lowest mean values, and the quadratic weighted kappa producing the highest mean values. The application of quadratic weights to the ordinally ranked traits (anterior nasal spine - ANS, inferior nasal margin - INA, malar tubercle - MT, nasal aperture width - NAW, and posterior zygomatic tubercle - PZT) consistently yielded higher agreement scores than if no weights were assigned. Closer inspection of the raw data revealed that this is because the scores for the ordinal ranked traits were nearly always within one score away. The binary traits (nasal overgrowth - NO, post-bregmatic depression - PBD) yielded the same kappa values regardless of weighting, as there are only two possible scores that can be assigned. While unranked traits with a greater number of trait states may yield scores that exhibit greater separation from the original score, often resulting in lower agreement values (i.e., an overestimation of error) if weights are assigned to them. Thus, using the correct weights that bests suits each of the different traits based on their data structure is highly recommended as it gives the most realistic results.

With the appropriate weights assigned to each trait, the intra-observer agreement ranged from 0.41 (moderate) to 1.00 (perfect), with nasal overgrowth (NO) and transverse palatine suture (TPS) performing the worst and best, respectively.

Table 4.1 – Comparison of intra-observer agreement using Cohen's kappa with different weights. Bold indicates values with moderate agreement or lower (<0.60).

	Unweighted kappa	Linear-weighted kappa	Quadratic-weighted kappa	Trait-specific mixed weights
ANS	0.62	0.72	0.82	0.82
INA	**0.47**	**0.52**	0.78	**0.47**
IOB	0.70	0.76	0.83	0.83
MT	**0.43**	**0.57**	0.72	0.72
NAS	0.62	0.62	0.62	0.62
NAW	0.84	0.87	0.91	0.91
NBC	0.64	0.75	0.84	0.64
NBS	**0.43**	0.63	0.79	**0.43**
NO	**0.41**	**0.41**	**0.41**	**0.41**
NFS	0.83	0.72	0.62	0.83
OS	0.80	0.84	0.89	0.80
PBD	0.74	0.74	0.74	0.74
PZT	**0.41**	**0.55**	0.69	0.69
SPS	0.81	0.72	0.64	0.81
TPS	1.00	1.00	1.00	1.00
PS	0.71	0.63	**0.56**	0.71
ZS	0.74	0.74	0.76	0.74
Mean	0.66	0.69	0.74	0.72
Min	0.41	0.41	0.41	0.41
Max	1.00	1.00	1.00	1.00

The inter-observer repeatability of the traits was compared among five observers with varying experience. This was done by comparing each observer to the primary observer (observer A), and then calculating the mean kappa value for each trait (Table 4.2). Overall, the mean kappa values ranged between -0.13 (poor) and 0.66 (substantial), with the nasal bone contour (NBC) performing the worst and interorbital breadth (IOB) performing the best. Interorbital breadth was the only trait to demonstrate substantial agreement, with all other traits showing moderate to poor repeatability. The performance of each observer compared to the primary observer revealed ariable results. For example, the anterior nasal spine (ANS) showed fair agreement between observers A and B (0.29) but showed almost perfect agreement between observers A and D (0.82). Conversely, orbit shape (OS) showed almost perfect agreement between observers A and B (0.83), while there was only slight agreement between observers A and D (0.15). Thus, each observer varied in which traits they were less/more repeatable.

In some instances, kappa values could not be calculated (e.g., NaN was obtained for nasal bone contour – NBC, post-bregmatic depression – PBD, and palate shape – PS). This indicates potential prevalence issues with the sample (one trait state is present in the sample and is being scored the most) which violates Cohen's kappa and prevents the calculation of a kappa value. However, since it only happened sporadically and did not happen with the intra-observer tests, it likely indicates that one observer in the pairwise comparisons were assigning the same score to all of the crania for the traits in question, while other observer pairs were assigning more variable scores; i.e., a bias issue rather than a prevalence issue.

Table 4.2 – Inter-observer agreement using Cohen's kappa among multiple observers. Scores recorded before any trait discussion. Bold indicates substantial agreement or higher (>0.61).

	Obs A – Obs B	**Obs A – Obs C**	**Obs A – Obs D**	**Obs A – Obs E**	**Mean**
ANS	0.29	0.42	**0.82**	**0.67**	0.55
INA	0.08	0.49	0.11	0.36	0.26
IOB	0.58	**0.74**	**0.91**	0.42	**0.66**
MT	0.55	**0.69**	0.55	0.35	0.53
NAS	0.48	-0.15	0.51	0.36	0.30
NAW	**0.66**	0.55	0.30	0.40	0.48
NBC	-0.09	-0.23	NaN	-0.09	-0.13
NBS	0.39	0.30	0.38	0.46	0.38
NO	0.05	-0.11	0.29	-0.11	0.03
NFS	0.40	**0.65**	0.47	0.41	0.48
OS	**0.83**	0.43	0.15	0.43	0.46
PBD	0.21	-0.32	NaN	0.05	-0.02
PZT	0.48	0.31	0.53	0.49	0.45
SPS	0.03	-0.08	0.34	0.61	0.23
TPS	0.57	0.21	0.29	0.09	0.29
PS	NaN	-0.33	0.43	0.37	0.16
ZS	**0.73**	0.52	0.52	**0.62**	0.60
Mean	0.39	0.24	0.44	0.35	0.34
Min	-0.09	-0.33	0.11	-0.11	-0.13
Max	0.83	0.74	0.91	0.67	0.66

All the observers rescored the same crania following a group discussion on the scoring procedure (Table 4.3). Overall, the mean kappa values increased after the discussion, ranging from -0.04 (poor) to 0.75 (substantial), with the supra-nasal suture (SPS) performing the worst and nasal aperture width (NAW) performing the best. Five traits demonstrated substantial

agreement values or higher (anterior nasal spine – ANS, interorbital breadth – IOB, nasal aperture width – NAW, nasal overgrowth – NO, and posterior zygomatic tubercle – PZT) compared to the first round of scores where only one trait demonstrated substantial agreement. Notably, four of the five traits with substantial agreement are ordinally ranked.

Mixed results were observed when comparing the mean kappa values for each observer. Even though observer B has five traits with substantial agreement, they presented with the overall lowest mean, indicating more variation in their scores. The mean kappa values decreased for both observers B and D after the discussion. For observer B the agreement remained fair, while with observer D the overall agreement dropped from moderate to fair. Both observers C and E showed increased agreement from fair to moderate after the trait discussion, with observer C demonstrating the most marked increase.

Table 4.3 – Inter-observer agreement using Cohen's kappa among multiple observers. Scores recorded after discussion session. Bold indicates substantial agreement or higher (>0.61).

	Obs A – Obs B	**Obs A – Obs C**	**Obs A – Obs D**	**Obs A – Obs E**	**Mean**
ANS	0.44	**0.66**	**1.00**	**0.64**	**0.69**
INA	0.23	**0.86**	-0.11	0.45	0.36
IOB	**0.77**	**0.91**	0.31	**0.77**	**0.69**
MT	0.44	0.59	0.59	0.58	0.55
NAS	**0.83**	0.24	0.41	-0.06	0.36
NAW	**0.81**	**0.91**	0.58	**0.72**	**0.75**
NBC	0.21	0.13	0.25	0.05	0.16
NBS	-0.06	0.44	0.55	0.26	0.30
NO	**0.80**	**0.78**	0.60	0.60	**0.70**
NFS	0.33	**0.67**	0.49	0.53	0.51
OS	0.39	0.57	**0.80**	0.09	0.46
PBD	-0.11	0.29	-0.15	**1.00**	0.26
PZT	0.21	**0.72**	**0.88**	**0.72**	**0.64**
SPS	0.17	0.11	-0.32	-0.11	-0.04
TPS	0.10	0.47	0.18	0.37	0.28
PS	**0.74**	0.18	0.55	0.28	0.44
ZS	0.11	**1.00**	0.06	0.33	0.37
Mean	0.38	0.56	0.39	0.42	0.44
Min	-0.11	0.11	-0.11	-0.11	-0.04
Max	0.83	0.91	1.00	0.77	0.75

4.2 Exploratory analyses: Trait frequencies and group differences

Table 4.4 presents the frequencies for the MMS traits. The sample size varies for each trait because of the presence of post-mortem damage, ante-mortem trauma, and tooth loss. A substantial amount of group overlap was observed for the traits, and not a single trait can be considered characteristic of a population (Figure 4.1). Kruskal-Wallis tests were used to identify potential population group differences (Table 4.5). Overall, 13 out of the 17 traits were noted to differ significantly among the population groups ($p<0.05$). The nasal bone shape (NBS), supra-nasal suture (SPS), transverse palatine suture (TPS), and palate shape (PS) were not significantly different. Since Kruskal-Wallis only indicates if there are any differences, a *post-hoc* Dunn's test was then used to further explore the variation among the three populations (see Table 4.6 for a breakdown of the group overlap). Five traits demonstrate no overlap (i.e., should be useful for distinguishing among all the groups); this includes the inferior nasal margin (INA), malar tubercle (MT), nasal aperture shape (NAS), nasal bone contour (NBC), and zygomaticomaxillary suture (ZS). The remainder of the traits demonstrated overlap between at least two of the groups. Black and coloured South Africans were observed to overlap more frequently, with some traits also presenting with overlap between coloured South Africans and white South Africans. However, none of the traits indicate significant overlap between black South Africans and white South Africans, suggesting the two groups are most dissimilar from each another. Ultimately, white South Africans more frequently presented with prominent anterior nasal spines (ANS), sharp inferior nasal margins (INA), narrow inter-orbital breadths (IOB), teardrop shaped nasal apertures (NAS), plateauing nasal bone contours (NBC), and nasal overgrowth (NO) compared to the other groups. On the other hand, black and coloured South Africans more frequently presented with small anterior nasal spines (ANS), rounded inferior nasal margins (INA), bowed nasal apertures (NAS), rounded nasal bone contours (NBC), and post-bregmatic depressions (PBD). However, while coloured South Africans overlapped with black South Africans, the coloured group more frequently yielded intermediate scores rather than extreme scores. While it was not within the scope of the study to specifically explore sex variation, Kruskal-Wallis tests did indicate that there were significant sex differences for seven of the traits (Table 4.5).

Table 4.4 – Trait frequencies for the three population groups. Refer to Table 3.2 for trait abbreviations.

	Population group					
	Black		Coloured		White	
Trait scores	n	%	n	%	n	%
ANS	(n = 220)		(n = 212)		(n= 207)	
1	143	65.0	115	54.2	25	12.1
2	66	30.0	85	40.1	79	38.2
3	11	5.0	12	5.7	103	49.7
INA	(n = 220)		(n = 219)		(n = 220)	
1	53	24.1	7	3.2	0	0.0
2	79	35.9	36	16.4	3	1.4
3	74	33.6	118	56.5	38	17.3
4	9	4.1	47	21.5	107	48.6
5	5	2.3	11	5.0	72	32.7
IOB	(n = 220)		(n = 219)		(n = 220)	
1	23	10.5	33	15.1	134	60.9
2	99	45.0	99	45.2	77	35.0
3	98	44.5	87	39.7	9	4.1
MT	(n = 218)		(n = 214)		(n = 220)	
0	2	1.0	0	0.0	16	7.3
1	116	53.2	151	70.6	167	75.9
2	75	34.4	59	27.6	34	15.5
3	25	11.5	4	1.9	3	1.4
NAS	(n = 220)		(n = 218)		(n = 220)	
1	28	12.7	65	29.8	183	83.2
2	36	16.4	17	7.8	28	12.7
3	156	70.9	136	62.4	9	4.1
NAW	(n = 220)		(n = 219)		(n = 220)	
1	5	2.3	6	2.7	80	36.4
2	67	30.5	74	33.8	113	51.4
3	148	67.3	139	63.5	27	12.2
NBC	(n = 194)		(n = 187)		(n = 202)	
0	116	59.8	70	37.4	0	0.0
1	44	22.7	87	46.5	39	19.3
2	7	3.6	7	3.7	79	39.1
3	9	4.6	14	7.5	78	38.6
4	18	9.3	9	4.8	6	3.0
NBS	(n = 213)		(n = 204)		(n = 214)	
1	58	27.2	25	12.3	32	15.0
2	107	50.2	153	75.4	167	78.0
3	26	12.2	7	3.4	12	5.6
4	22	10.3	18	8.9	3	1.4
NO	(n = 208)		(n = 186)		(n = 205)	
0	202	97.1	186	100.0	168	82.0
1	6	2.9	0	0.0	37	18.0
NFS	(n = 202)		(n = 200)		(n = 214)	
1	73	36.1	96	48.0	123	57.5

2	71	35.1	58	29.0	38	17.8
3	17	8.4	16	8.0	23	10.7
4	41	20.3	30	15.0	30	14.0
OS	(n = 219)		(n = 218)		(n = 220)	
1	118	53.9	159	72.9	150	68.2
2	89	40.6	44	20.2	49	22.3
3	12	5.5	15	6.9	21	9.5
PBD	(n = 218)		(n = 214)		(n = 217)	
0	144	65.1	155	72.4	176	81.1
1	74	33.9	59	27.6	41	18.9
PZT	(n = 218)		(n = 217)		(n = 220)	
0	14	6.4	6	2.8	25	11.4
1	77	35.3	65	30.0	104	47.3
2	72	33.0	91	41.9	63	28.6
3	55	25.2	55	25.3	28	12.7
SPS	(n = 219)		(n = 220)		(n = 220)	
0	69	31.5	29	13.2	23	10.5
1	19	8.7	85	38.6	89	40.5
2	131	59.8	106	48.2	108	49.0
TPS	(n = 213)		(n = 211)		(n = 215)	
1	53	24.9	54	25.6	59	27.4
2	110	51.6	119	56.4	126	58.6
3	23	10.8	15	7.1	14	6.5
4	27	12.7	23	10.9	16	7.5
PS	(n = 168)		(n = 116)		(n = 53)	
1	50	29.8	31	26.7	25	47.2
2	29	17.3	18	15.5	9	17.0
3	54	32.1	55	47.4	11	20.8
4	35	20.8	12	10.3	8	15.1
ZS	(n = 210)		(n = 209)		(n = 215)	
0	153	72.9	84	40.2	75	34.9
1	45	21.4	123	58.8	112	52.1
2	12	5.7	2	1.0	28	13.0

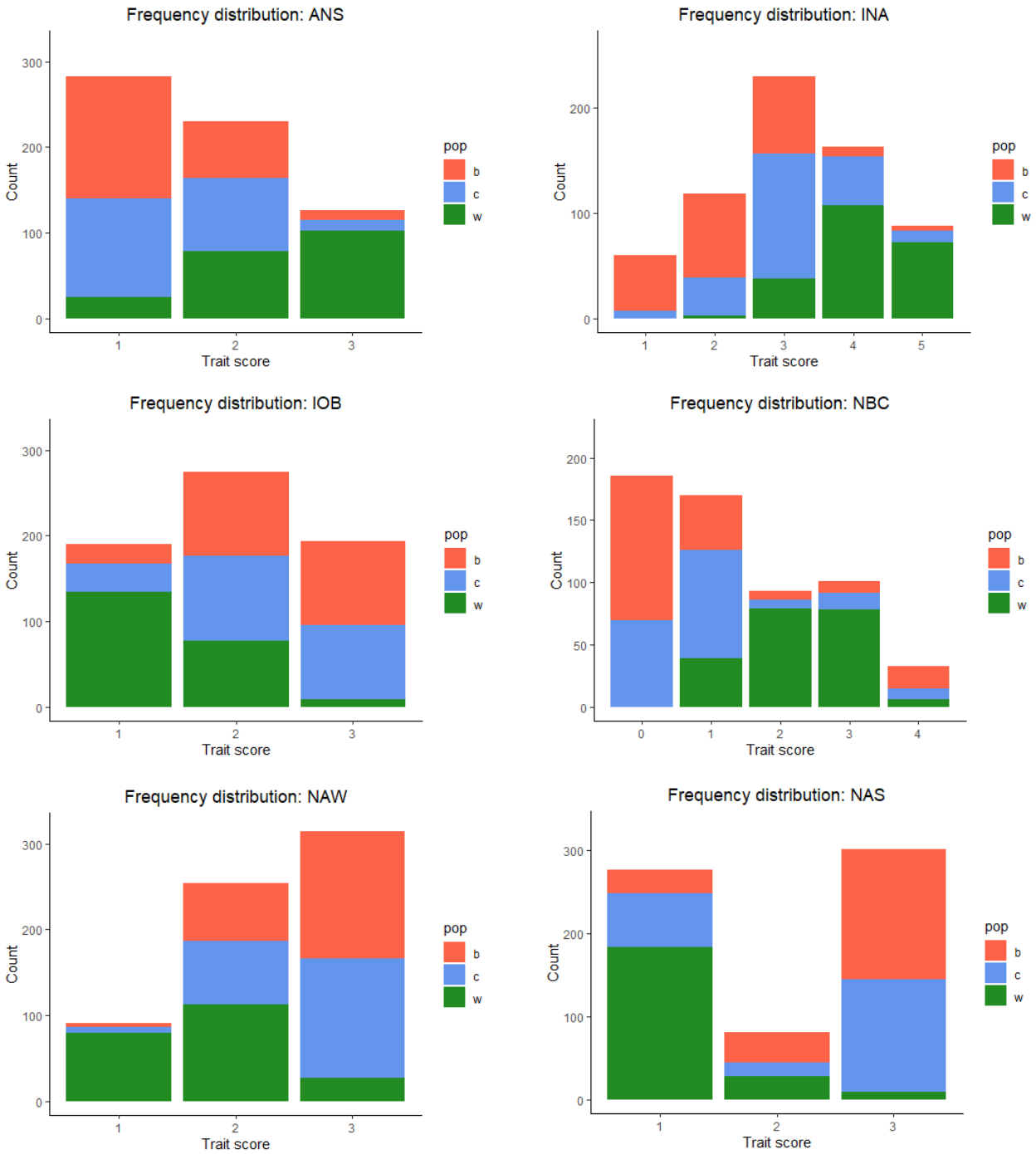


Figure 4.1 – Frequency distribution for a selection of traits to demonstrate group variation and overlap.

Table 4.5 - Results of the Kruskal-Wallis test comparing trait score frequencies among the populations and between the sexes. Bold indicates significant differences.

Trait	Population	Sex
ANS	**<0.05**	0.08
INA	**<0.05**	**<0.05**
IOB	**<0.05**	**<0.05**
MT	**<0.05**	**<0.05**
NAS	**<0.05**	0.76
NAW	**<0.05**	**<0.05**
NBC	**<0.05**	0.05
NBS	0.28	0.24
NO	**<0.05**	0.33
NFS	**<0.05**	0.33
OS	**<0.05**	0.18
PBD	**<0.05**	0.07
PZT	**<0.05**	**<0.05**
SPS	0.92	**<0.05**
TPS	0.19	0.93
PS	0.06	**<0.05**
ZS	**<0.05**	0.99

Table 4.6 - Break down of group overlap for trait scores based on the Kruskal-Wallis and Dunn's tests.

No groups overlap	All groups overlap	B and C overlap	B and W overlap	W and C overlap
INA	NBS	ANS	-	NFS
MT	SPS	IOB		OS
NAS	TPS	NAW		PBD
NBC	PS	NO		
ZS		PBD		
		PZT		

4.3 Correlations

Polychoric correlations were used to analyse the MMS traits, with correlation coefficients ranging between -0.57 and 0.60 (Table 4.7). Most of the notable correlations were observed for features located in the nasal and orbital region. Fair positive correlations were noted between the nasal aperture shape (NAS) and the inter-orbital breadth (IOB) (r = 0.59), nasal aperture shape (NAS) and nasal width (NAW) (r = 0.59), as well as the anterior nasal spine (ANS) and inferior nasal margin (INA) (r = 0.57). Additionally, the nasal aperture width and inter-orbital breadth (r = 0.60) were noted to be moderately correlated.

Table 4.7 – Polychoric correlations demonstrating the relationship between macromorphoscopic traits.

	ANS	INA	IOB	MT	NAS	NAW	NBC	NBS	NO	NFS	OS	PBD	PZT	SPS	TPS	PS	ZS
ANS	-																
INA	0.57	-															
IOB	-0.35	-0.45	-														
MT	-0.19	-0.27	0.25	-													
NAS	-0.50	-0.57	0.59	0.27	-												
NAW	-0.46	-0.53	0.60	0.22	0.59	-											
NBC	0.43	0.38	-0.44	-0.14	-0.52	-0.47	-										
NBS	-0.03	-0.19	-0.06	0.01	0.08	0.09	0.03	-									
NO	0.27	0.31	-0.55	-0.23	-0.55	-0.37	0.43	-0.13	-								
NFS	-0.01	-0.08	0.09	0.09	0.10	0.04	-0.13	-0.06	-0.04	-							
OS	-0.02	-0.02	-0.14	0.21	-0.01	-0.06	-0.01	0.01	0.14	0.09	-						
PBD	-0.11	-0.10	0.09	0.06	0.09	0.09	-0.10	0.06	-0.01	-0.06	0.02	-					
PZT	-0.13	-0.21	0.26	0.16	0.16	0.21	-0.18	0.08	-0.03	-0.01	-0.12	0.12	-				
SPS	0.09	-0.01	0.01	0.04	-0.02	-0.04	0.09	0.09	0.13	0.07	-0.05	-0.02	0.07	-			
TPS	-0.05	-0.08	0.09	-0.03	0.06	0.04	-0.04	-0.04	0.01	0.02	-0.04	-0.09	0.02	-0.02	-		
PS	-0.03	-0.19	0.15	0.12	0.14	0.22	-0.05	0.03	0.02	0.11	-0.02	-0.02	-0.02	0.14	0.12	-	
ZS	0.18	0.21	-0.20	-0.12	-0.26	-0.16	0.22	0.03	0.16	-0.06	-0.07	-0.07	-0.06	0.04	0.03	0.08	-

CHAPTER 5: RESULTS - CRANIOMETRIC VARIATION

5.1 Observer agreement

Technical error of measurement was used to gauge the degree of measurement repeatability between two observers (Table 5.1). The intra-observer TEM and %TEM ranged between 0.32 and 0.74, and 0.30% and 2.46%, respectively. The margin of error was slightly higher for the inter-observer analysis, with the TEM and %TEM ranging between 0.45 and 1.67, and 0.34% and 5.71%, respectively. The variables that presented with the greatest error for both inter- and intra-observer analyses were inter-orbital breadth and mastoid height. The dimensions of the palate (MAL and MAB) and foramen magnum (FOL and FOB) also demonstrated greater levels of measurement error for the inter-observer analysis.

Bland-Altman plots were used to illustrate the variability in the repeatability of the measurements. The intra-observer plot demonstrated greater precision, with the majority of the variables presenting with less than 2mm difference (Figure 5.1). The inter-observer plot demonstrated greater variation, with five measurements presenting with as much as 4mm difference (Figure 5.2).

Despite some observer variation, the combined TEM and Bland-Altman results demonstrate satisfactory levels of agreement, and all measurements were retained in the analyses.

Table 5.1 - Absolute technical error of measurement (TEM) and relative technical error of measurement (%TEM) for inter- and intra-observer agreement.

	Intra-observer error		Inter-observer error	
	TEM	**%TEM**	**TEM**	**%TEM**
GOL	0.62	0.33	0.81	0.43
XCB	1.02	0.78	1.07	0.81
ZYB	0.39	0.30	0.59	0.46
BBH	0.45	0.34	0.45	0.34
BNL	0.39	0.38	0.45	0.44
BPL	0.47	0.47	0.82	0.81
MAL	0.78	1.37	1.67	2.93
MAB	0.76	1.18	1.48	2.29
ASB	0.67	0.62	0.55	0.51
NPH	0.91	1.35	0.71	1.05
WFB	0.39	0.40	0.59	0.61
UFBR	0.32	0.30	0.55	0.52
NLH	0.63	1.27	0.84	1.68
NLB	0.39	1.51	0.45	1.75
OBB	0.71	1.77	0.59	1.50
OBH	0.32	0.95	0.59	1.78
EKB	0.89	0.92	0.81	0.82
DKB	0.63	2.46	1.47	5.71
FRC	0.67	0.59	0.74	0.66
PAC	0.91	0.79	1.13	0.97
OCC	0.67	0.69	0.62	0.64
FOL	0.22	0.62	0.89	2.48
FOB	0.50	1.71	0.63	2.17
MDH	0.74	2.45	1.00	3.30
AUB	0.50	0.43	0.55	0.47
Mean	0.60	0.96	0.80	1.40
Min	0.32	0.30	0.45	0.34
Max	0.74	2.46	1.67	5.71

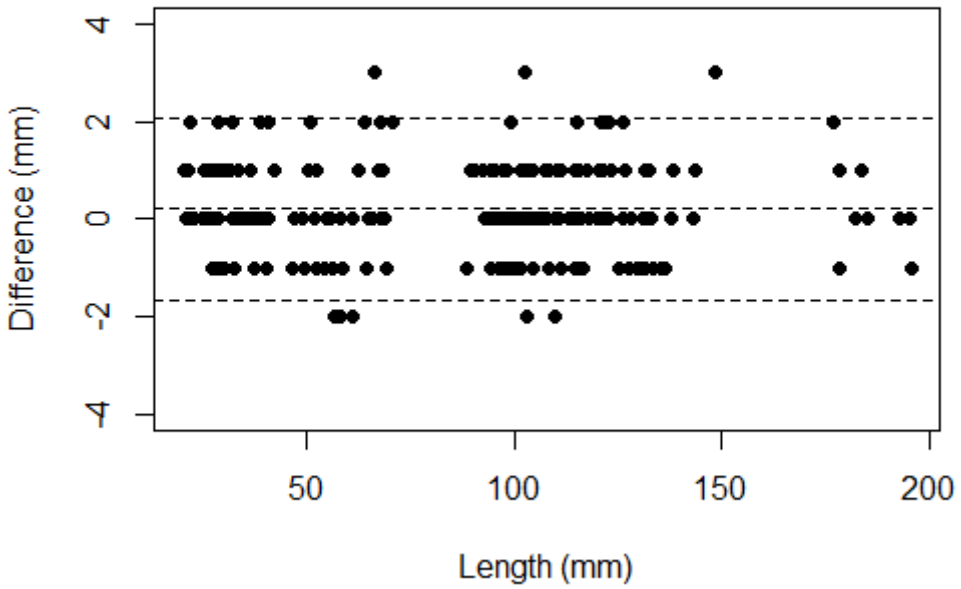


Figure 5.1 – Bland-Altman plot illustrating the intra-observer agreement of measurements.

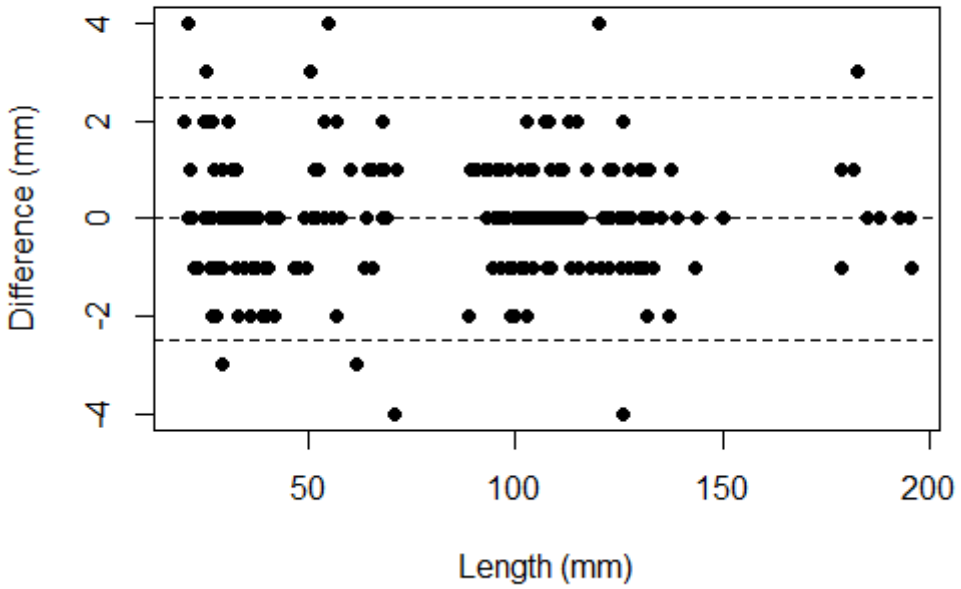


Figure 5.2 – Bland-Altman plot illustrating the inter-observer agreement of measurements compared between two observers.

5.2 Exploratory analyses: Measurement means and group differences

Table 5.2 presents the measurement means and standard deviations for each population group. Overall, a lot of variation was observed with no trend in terms of which group is consistently the largest or smallest for any particular region. All measurements were noted to have significant differences in the ANOVAs for ancestry, except for the foramen magnum length (FOL) (Table 5.3). Further analysis with Tukey's HSD revealed that 10 out of the 25 measurements were significantly different among all three groups; this includes several variables pertaining to facial and cranial breadth (Table 5.4). The remainder of the measurements demonstrated overlap between at least two of the three groups, with no particular trends regarding cranial regions. When considering sex, all measurements were noted to differ significantly between males and females, except for orbital height (OBH).

Table 5.2 – Summary statistics showing the measurement means (mm) and standard deviations for black, white, and coloured South Africans. Refer to Table 3.3 for measurement abbreviations.

	Black			White			Coloured		
Variable	***n***	**mean**	**sd**	***n***	**mean**	**sd**	***n***	**mean**	**sd**
GOL	217	183.9	7.4	200	182.2	7.5	206	181.0	8.7
XCB	217	131.0	5.3	212	135.5	5.7	209	132.9	5.5
ZYB	200	126.5	6.3	213	124.6	5.9	157	123.5	6.8
BBH	213	130.5	6.1	202	133.6	6.2	206	128.6	5.8
BNL	219	99.4	5.1	219	99.6	5.2	218	97.0	5.1
BPL	167	101.0	5.9	96	93.5	5.1	114	95.5	6.1
MAL	166	56.3	4.0	87	53.1	3.4	112	53.0	4.2
MAB	128	63.5	3.9	14	60.0	4.6	28	60.0	4.1
ASB	211	106.1	5.0	208	112.7	4.9	215	107.4	5.1
NPH	167	65.7	4.6	95	67.2	3.9	114	63.6	5.3
WFB	218	96.3	4.8	218	95.4	4.2	217	94.1	5.1
UFBR	217	105.8	4.7	220	102.2	4.5	202	103.4	4.7
NLH	220	47.6	3.0	219	50.6	3.2	212	47.2	3.3
NLB	215	27.4	2.2	218	23.4	2.3	209	26.3	2.5
OBB	219	39.6	2.0	220	40.1	2.0	216	39.6	2.1
OBH	220	33.7	2.0	220	34.1	1.8	216	33.6	2.5
EKB	218	98.7	4.4	219	95.3	4.2	207	96.5	4.4
DKB	216	25.1	2.8	217	21.5	2.7	216	22.8	2.7
FRC	213	112.1	5.6	203	112.4	5.4	208	110.4	5.7
PAC	212	114.1	6.5	194	112.8	6.5	204	112.0	6.8
OCC	212	96.1	5.2	207	98.6	5.1	205	94.8	5.9
FOL	218	37.0	2.9	219	37.4	2.8	218	36.9	2.9
FOB	219	29.1	2.4	219	30.6	2.5	220	28.8	2.2
MDH	218	27.9	3.4	220	29.9	3.2	219	26.2	3.6
AUB	218	115.9	5.3	218	119.7	5.0	219	114.6	7.7

Table 5.3 – ANOVA and Kruskal-Wallis results evaluating the effects of population and sex for each measurement. Bold indicates significant.

	Population		Sex	
Variable	**F-value**	**Pr (< F)**	**F-value**	**Pr (< F)**
GOL	7.48	**<0.01**	259.10	**<0.01**
XCB	35.63	**<0.01**	54.27	**<0.01**
ZYB	10.78	**<0.01**	301.0	**<0.01**
BBH	35.58	**<0.01**	172.80	**<0.01**
BNL	18.38	**<0.01**	196.50	**<0.01**
BPL	60.98	**<0.01**	46.84	**<0.01**
MAL	32.13	**<0.01**	48.92	**<0.01**
MAB	12.48	**<0.01**	36.66	**<0.01**
ASB	103.10	**<0.01**	35.76	**<0.01**
NPH	15.58	**<0.01**	57.48	**<0.01**
WFB	12.74	**<0.01**	45.60	**<0.01**
UFBR	33.32	**<0.01**	128.60	**<0.01**
NLH	73.86	**<0.01**	80.22	**<0.01**
NLB	133.80	**<0.01**	42.79	**<0.01**
OBB	4.15	**0.02**	76.67	**<0.01**
OBH	3.54	**0.03**	3.35	0.07
EKB	34.99	**<0.01**	111.30	**<0.01**
DKB	95.51	**<0.01**	7.17	**<0.01**
FRC	8.25	**<0.01**	135.40	**<0.01**
PAC	5.36	**0.01**	72.77	**<0.01**
OCC	26.12	**<0.01**	14.0	**<0.01**
FOL	1.49	0.23	84.09	**<0.01**
FOB	37.08	**<0.01**	57.18	**<0.01**
MDH	63.92	**<0.01**	78.40	**<0.01**
AUB	40.76	**<0.01**	100.60	**<0.01**

Table 5.4 – Break down of group overlap for measurement means based on the ANOVA and Tukey's HSD tests.

No groups overlap	All groups overlap	B and C overlap	B and W overlap	W and C overlap
ASB	FOL	AUB	BNL	GOL
BBH		FOB	FRC	MAB
BPL		NLH	GOL	MAL
DKB		OBB	NPH	OBB
EKB		OBH	OBH	PAC
MDH			PAC	ZYB
NLB			WFB	
OCC				
UFBR				
XCB				

5.3 Correlations

Pearson correlations were used to analyse the cranial measurements. Overall, much stronger correlations were observed among the measurements compared to the MMS traits, with correlation coefficients ranging between -0.12 and 0.92. Very strong positive correlations were noted between the bizygomatic breadth (ZYB) and the upper facial breadth (UFBR) (r = 0.82) and auricular breadth (AUB) (r = 0.84), respectively; upper facial breadth (UFBR) was also strongly correlated with the bi-orbital breadth (EKB) (r = 0.92). Similarly, the palate length (MAL) and basion-prosthion length (BPL) were highly correlated (0.81). Numerous positive correlations of moderate strength (r > 0.6) were also recorded across the entire cranium. Few negative correlations were observed with the measurements; this includes orbital height (OBH) and inter-orbital breadth (DKB) (r = -0.12), as well as the occipital cord (OCC) with the inter-orbital breadth (DKB) (r = -0.08), parietal cord (PAC) (r = -0.05), and basion prosthion length (BPL)(r = -0.03). However, the negative correlations were very weak.

Finally, polyserial correlations were used to analyse the relationship between the MMS traits and cranial measurements. Moderate positive correlations were noted between the nasal breadth measurement (NLB) and the inter-orbital breadth (IOB) (r = 0.55), nasal aperture shape (NAS) (0.50) and nasal aperture width (NAW) scores (r = 0.71). The inter-orbital breadth (IOB) score was also moderately correlated with both the bi-orbital (EKB) (r = 0.59) and inter-orbital

breadth (DKB) measurements (r = 0.60). The presence of correlations among the variables collected from the same region (i.e., the face) is not unexpected.

Table 5.5 – Pearson correlations demonstrating the relationship between the cranial measurements.

	GOL	XCB	ZYB	BBH	BNL	BPL	MAL	MAB	ASB	NPH	WFB	UFBR	NLH	NLB	OBB	OBH	EKB	DKB	FRC	PAC	OCC	FOL	FOB	MDH	AUB
GOL	-																								
XCB	0.41	-																							
ZYB	0.64	0.44	-																						
BBH	0.56	0.42	0.43	-																					
BNL	0.69	0.26	0.55	0.64	-																				
BPL	0.57	0.04	0.56	0.34	0.69	-																			
MAL	0.52	0.08	0.54	0.36	0.49	0.81	-																		
MAB	0.58	0.31	0.69	0.38	0.47	0.52	0.56	-																	
ASB	0.42	0.61	0.38	0.41	0.29	0.08	0.11	0.25	-																
NPH	0.50	0.27	0.31	0.44	0.46	0.38	0.40	0.38	0.22	-															
WFB	0.51	0.49	0.60	0.48	0.43	0.24	0.24	0.48	0.40	0.16	-														
UFBR	0.63	0.41	0.82	0.42	0.56	0.52	0.47	0.66	0.35	0.22	0.76	-													
NLH	0.41	0.36	0.33	0.30	0.41	0.13	0.22	0.25	0.39	0.72	0.25	0.25	-												
NLB	0.53	0.18	0.56	0.17	0.40	0.47	0.35	0.53	0.10	0.04	0.35	0.60	0.08	-											
OBB	0.46	0.35	0.60	0.31	0.46	0.35	0.28	0.40	0.25	0.12	0.44	0.63	0.20	0.44	-										
OBH	0.15	0.18	0.17	0.07	0.10	0.08	0.10	0.03	0.12	0.33	0.14	0.16	0.39	0.12	0.28	-									
EKB	0.65	0.34	0.79	0.37	0.53	0.51	0.45	0.63	0.30	0.15	0.67	0.92	0.15	0.68	0.70	0.14	-								
DKB	0.35	0.06	0.44	0.16	0.31	0.32	0.26	0.43	0.08	0.01	0.52	0.60	0.02	0.43	0.03	-0.12	0.56	-							
FRC	0.68	0.53	0.47	0.70	0.49	0.39	0.38	0.48	0.45	0.49	0.48	0.46	0.33	0.30	0.30	0.15	0.43	0.24	-						
PAC	0.66	0.13	0.36	0.52	0.43	0.35	0.29	0.36	0.23	0.22	0.42	0.40	0.14	0.37	0.31	0.04	0.43	0.36	0.48	-					
OCC	0.36	0.33	0.13	0.30	0.06	-0.03	0.13	0.08	0.28	0.27	0.11	0.09	0.26	0.04	0.02	0.07	0.07	-0.08	0.20	-0.05	-				
FOL	0.42	0.19	0.47	0.28	0.29	0.26	0.26	0.28	0.28	0.27	0.27	0.43	0.27	0.29	0.37	0.22	0.43	0.21	0.30	0.20	0.09	-			
FOB	0.31	0.27	0.39	0.24	0.27	0.23	0.17	0.14	0.28	0.27	0.19	0.25	0.30	0.12	0.25	0.19	0.23	0.05	0.26	0.09	0.18	0.54	-		
MDH	0.39	0.28	0.43	0.35	0.27	0.21	0.35	0.38	0.33	0.24	0.30	0.39	0.27	0.20	0.32	0.14	0.36	0.15	0.39	0.22	0.16	0.25	0.19	-	
AUB	0.56	0.64	0.84	0.45	0.45	0.29	0.32	0.55	0.56	0.32	0.60	0.67	0.45	0.35	0.52	0.15	0.64	0.25	0.51	0.29	0.19	0.37	0.35	0.45	-

Table 5.6 – Polyserial correlations demonstrating the relationship between macromorphoscopic traits and cranial measurements.

	ANS	INA	IOB	MT	NAS	NAW	NBC	NBS	NO	NFS	OS	PBD	PZT	SPS	TPS	PS	ZS
GOL	0.05	-0.14	0.26	0.11	0.07	0.15	-0.03	-0.01	-0.04	-0.05	-0.08	-0.03	0.11	0.15	0.05	0.16	-0.02
XCB	0.21	0.21	-0.02	-0.01	-0.21	-0.12	0.24	-0.08	0.07	-0.10	-0.02	-0.01	-0.04	0.10	-0.03	-0.04	0.16
ZYB	0.01	-0.18	0.34	0.14	0.11	0.19	0.01	-0.04	-0.05	0.01	-0.09	-0.01	0.25	0.12	0.06	0.22	-0.05
BBH	0.22	0.09	0.010	-0.03	-0.13	-0.09	0.14	-0.08	0.05	-0.12	-0.03	-0.13	0.02	0.17	0.01	0.12	0.10
BNL	0.12	-0.08	0.17	0.02	0.03	0.08	-0.01	-0.01	-0.01	-0.12	-0.10	-0.04	0.15	0.20	-0.01	0.14	-0.01
BPL	-0.20	-0.46	0.36	0.18	0.34	0.32	-0.21	0.13	-0.11	-0.14	-0.07	0.01	0.16	0.16	0.02	0.27	-0.15
MAL	-0.16	-0.36	0.28	0.10	0.18	0.19	-0.10	0.05	-0.06	0.11	-0.11	-0.01	0.13	0.16	0.03	0.29	-0.10
MAB	-0.10	-0.26	0.40	0.19	0.16	0.22	-0.01	-0.19	0.06	0.08	-0.11	-0.03	0.23	0.16	0.10	0.26	-0.03
ASB	0.33	0.30	-0.14	-0.08	-0.34	-0.24	0.30	-0.09	0.17	0.05	-0.06	-0.11	-0.06	0.10	-0.07	-0.05	0.20
NPH	0.23	0.05	-0.02	-0.04	-0.13	-0.16	0.17	-0.01	0.05	-0.06	-0.06	-0.05	0.06	0.20	-0.04	0.11	0.05
WFB	-0.01	-0.07	0.38	0.08	0.07	0.16	-0.05	-0.06	-0.03	-0.01	-0.13	0.05	0.07	0.09	-0.03	0.02	-0.05
UFBR	-0.14	-0.29	0.56	0.18	0.26	0.36	-0.20	-0.03	-0.12	-0.10	-0.19	0.08	0.22	0.12	0.01	0.16	-0.12
NLH	0.250	0.22	-0.14	-0.11	-0.28	-0.16	0.27	-0.07	0.09	-0.02	-0.03	-0.04	0.03	0.10	-0.06	-0.01	0.10
NLB	-0.36	-0.53	0.55	0.20	0.50	0.71	-0.37	0.02	-0.18	-0.03	-0.05	0.03	0.19	0.03	0.05	0.18	-0.16
OBB	0.13	0.09	0.18	0.06	-0.05	0.03	0.02	-0.01	0.01	0.03	-0.13	-0.02	0.14	0.08	-0.01	0.08	0.03
OBH	0.16	0.09	-0.06	0.04	-0.07	-0.08	0.14	-0.03	0.06	0.04	0.18	0.06	-0.01	0.04	-0.04	-0.02	0.09
EKB	-0.13	-0.29	0.59	0.20	0.27	0.37	-0.18	-0.02	-0.12	0.03	-0.16	0.02	0.21	0.08	0.04	0.19	-0.09
DKB	-0.25	-0.39	0.60	0.17	0.35	0.41	-0.29	-0.07	-0.14	0.01	-0.06	0.05	0.16	-0.02	0.04	0.05	-0.16
FRC	0.09	-0.04	0.10	0.02	0.01	0.08	0.06	-0.06	0.01	-0.01	-0.02	-0.05	-0.03	0.11	-0.03	0.14	0.07
PAC	-0.01	-0.10	0.13	0.12	0.01	0.09	-0.07	-0.03	0.02	-0.07	0.03	-0.02	0.08	0.09	0.09	0.08	-0.03
OCC	0.17	0.12	-0.08	-0.03	-0.14	-0.13	0.18	-0.10	0.06	-0.04	0.04	-0.01	-0.15	0.08	-0.01	-0.03	0.08
FOL	0.05	-0.01	0.09	-0.02	-0.03	0.04	0.12	-0.03	0.03	0.01	-0.04	-0.04	0.08	0.11	0.01	0.15	0.06
FOB	0.16	0.16	-0.08	-0.15	-0.16	-0.14	0.25	-0.05	0.13	-0.02	-0.05	-0.07	0.04	0.07	-0.04	0.00	0.15
MDH	0.25	0.16	-0.11	-0.03	-0.16	-0.13	0.19	-0.07	0.12	-0.03	-0.04	-0.06	0.01	0.08	-0.02	0.04	0.11
AUB	0.25	0.14	0.06	-0.05	-0.20	-0.10	0.24	-0.08	0.08	-0.04	-0.10	-0.04	0.10	0.11	-0.01	0.10	0.09

CHAPTER 6: RESULTS - RANDOM FOREST MODELS FOR CLASSIFICATION

6.1 MMS models

Univariate random forest models were first created to explore the predictive ability to estimate ancestry of each MMS trait on its own prior to building a multivariate model (Table 6.1). Palate shape (PS) was removed from any subsequent analyses as there were too many missing values resulting in a small sample.

Training accuracies for the univariate models ranged from 33.1% to 68.7%, with the transverse palatine suture (TPS) and nasal bone contour (NBC) performing the worst and best, respectively. The testing accuracies were comparable and ranged from 29.1% to 69.7%. Although testing accuracies are typically lower compared to training accuracies, several traits demonstrated an increased accuracy (up to 6% higher) when employed with an independent sample. Conversely, the kappa values yielded much lower accuracies, ranging from -0.06% to 54.6%. This indicates that many of the correct classifications achieved with the univariate MMS traits occurred because of chance rather than the model producing a true positive prediction.

Table 6.1 – Univariate classification accuracy (%) of each MMS trait using RFM for population affinity.

Trait	Training accuracy	Testing accuracy	Kappa
ANS	52.3	46.7	20.0
INA	66.3	61.8	42.7
IOB	46.1	52.7	42.8
MT	42.8	43.6	15.5
NAS	57.4	55.2	32.7
NAW	50.9	53.9	30.9
NBC	68.7	69.7	54.6
NBS	43.8	37.6	6.4
NO	39.0	38.8	8.2
NFS	41.6	46.1	19.1
OS	40.8	41.8	12.7
PBD	38.0	39.4	9.1
PZT	41.0	45.5	18.2
SPS	43.8	44.2	16.4
TPS	33.1	29.1	-0.06
ZS	49.5	48.5	22.7

All the MMS traits were then combined into a multivariate model. Overall, the MMS traits yielded an accuracy of 78.7%. Table 6.2 presents the training accuracies, with a breakdown of the predictive performance of each population group and group overlap. The greatest overlap (and subsequent misclassification) was observed between black and coloured South Africans. White South Africans had the least overlap, resulting in the highest group accuracy (89.7%). The model is not overfit, as indicated by the comparable, and slightly higher, testing accuracy of 81.8%. The kappa value was lower than both the training and testing accuracies (72.7%); however, the discrepancy is much smaller than observed with the univariate models and is still at a level that indicates good performance. This indicates that the multivariate model is less prone to produce false positive predictions resulting from chance than the univariate models.

Table 6.2 – Confusion matrix showing patterns of overlap and misclassification among the groups for the training model employing the MMS traits.

		Classifies into:			%
		Black	White	Coloured	Correct
Group:	Black	127	5	33	77.0
	White	3	148	14	89.7
	Coloured	32	18	115	69.7
				Total:	78.7

Finally, the variable importance was calculated to assess how much discriminatory power each trait contributes to the model and overall correct classification. Ultimately all the traits contributed some information to the model, with the variable importance ranging from 2.7 to 56.0 (Table 6.3). Figure 1 graphically demonstrates the contribution of each trait to the model. The highest ranked traits include the inferior nasal margin (INA), nasal bone contour (NBC), and nasal aperture shape (NAS) – i.e., variables in the nasal region. The lowest ranked traits include nasal overgrowth (NO), post-bregmatic depression (PBD), and orbit shape (OS).

Table 6.3 – RFM variable importance for MMS traits.

Trait	Variable importance
INA	56.0
NBC	50.0
NAS	33.8
ANS	23.3
ZS	19.9
IOB	19.6
NAW	16.2
SPS	15.9
PZT	14.7
NBS	14.4
MT	13.3
NFS	12.8
TPS	12.7
OS	10.7
PBD	6.9
NO	2.7

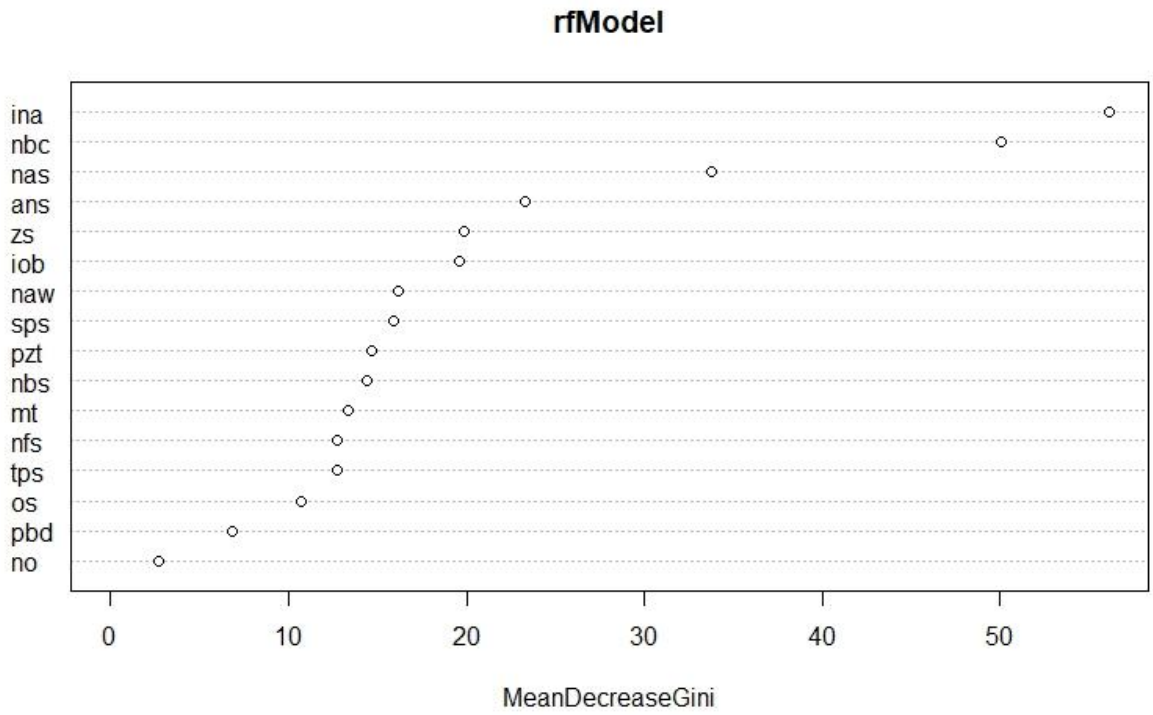


Figure 6.1 – Variable importance for the multivariate model employing all MMS traits.

An additional multivariate model was created to further explore the accuracy of the traits. For this model the number of traits employed were reduced. More specifically, all traits with poor repeatability as noted with Cohen's kappa, any trait that did not yield significant differences with Kruskal-Wallis, and any trait with low variable importance (< 15) were removed. Using seven traits, the model yielded training and testing accuracies of 73.6% and 75.3%, respectively, with a kappa value of 62.9%. Thus, a reduction in the number of traits included in the model led to a slight decrease in the model accuracy.

6.2 Craniometric models

Because of a large number of missing values, all measurements with landmarks surrounding the palate were removed from subsequent analyses (basion-prosthion length, palate length, palate breadth, nasion-prosthion height). Training accuracies for the univariate models ranged from 30.9% to 53.7%, with the orbital height (OBH) and nasal breadth (NLB) performing the worst and best, respectively (Table 6.4). Once again several measurements demonstrated an increase (up to 7.2% higher) with the testing accuracies, where the correct classifications ranged from 29.7% to 52.1%. The kappa values were lower than both the training and testing accuracies (between -5.5% and 31.8%).

Table 6.4 – Univariate classification accuracy (%) of each measurement using RFM for population affinity.

Trait	Training accuracy	Testing accuracy	Kappa
GOL	36.4	43.6	15.4
XCB	43.4	37.0	5.5
ZYB	45.1	49.1	23.6
BBH	39.0	41.2	11.8
BNL	35.2	37.6	6.4
ASB	46.1	45.5	18.2
WFB	42.6	29.7	-5.5
UFBR	42.2	38.8	8.2
NLH	43.0	42.4	13.6
NLB	53.7	52.1	28.2
OBB	33.8	35.2	2.7
OBH	30.9	37.0	5.5
EKB	43.4	45.5	18.2
DKB	52.1	54.6	31.8
FRC	35.8	38.2	7.3
PAC	39.8	42.2	13.6
OCC	44.2	41.2	4.4
FOL	32.5	29.7	-5.5
FOB	39.6	36.4	4.6
MDH	47.9	46.1	19.1
AUB	40.4	38.2	7.3

The overall accuracy for the multivariate measurement model was 72.3%; this is lower than the MMS trait model (6.4% decrease), but it does outperform all univariate models. Similar patterns of variation and therefore misclassifications were observed with the measurement data, such that black and coloured South Africans overlap more frequently, and white South Africans are more dissimilar (Table 6.5).

Table 6.5 – Confusion matrix showing patterns of overlap and misclassification among the groups for the training model employing the measurements.

		Classifies into:			%
		B	W	C	Correct
Group:	B	111	6	48	67.3
	W	6	147	12	89.1
	C	44	21	100	60.6
				Total:	72.3

The variable importance for the measurements were more evenly distributed than the MMS traits, ranging between 8.2 and 35.2 (Table 6.6). Thus, all the measurements contribute some information to the classification model. Figure 6.2 shows the measurement contributions to the model. The highest ranked measurements include the nasal breadth (NLB), biasterionic breadth (ASB), and inter-orbital breadth (DKB). The lowest ranked traits include orbital breadth (OBB), orbital height (OBH), and foramen magnum length (FOL).

Table 6.6 – RFM variable importance for measurements.

Trait	Variable importance
NLB	35.2
ASB	30.9
DKB	24.8
NLH	22.2
AUB	17.4
MDH	17.1
OCC	16.8
XCB	16.8
BBH	15.2
EKB	14.7
UFBR	13.7
ZYB	12.9
FOB	12.4
BNL	11.7
GOL	10.7
WFB	10.6
PAC	10.5
FRC	9.9
FOL	9.0
OBH	8.7
OBB	8.2

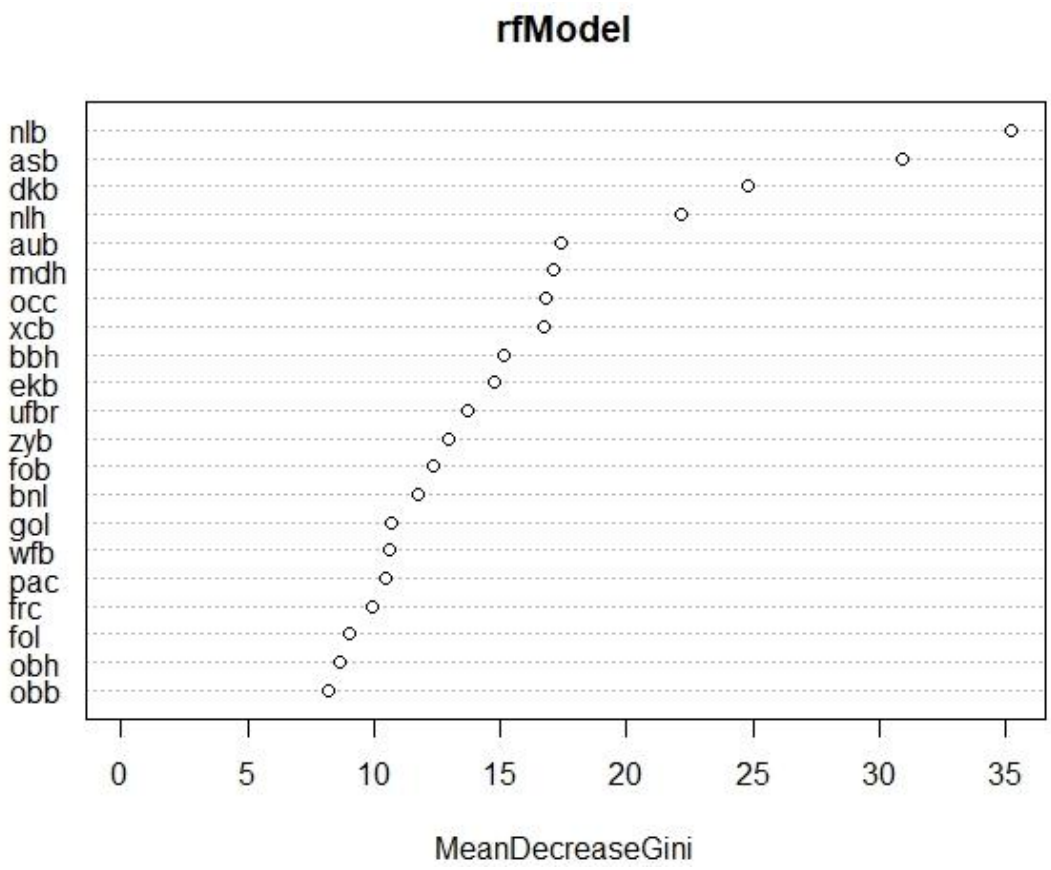


Figure 6.2 – Variable importance for the multivariate model employing all measurements.

An additional model was created, where all measurements with a variable importance below 14 were removed. Using 11 traits, the model yielded training and testing accuracies of 70.3% and 72.7%, respectively, with a kappa value of 59.1%. Once again, a reduction in the number of measurements included in the model led to a decrease in the model accuracy.

6.3 Mixed model

The MMS traits and measurements were combined to create a mixed multivariate model and the overall accuracy for the combined model was 81.0% (Table 6.7). The combined model achieved a classification accuracy higher than the accuracies achieved when using the MMS traits and measurements independently. Similar patterns of misclassification were observed with the combined model. Interestingly, the coloured South African group achieved a higher correct classification with the combined model than the separate craniometric and MMS multivariate models. More specifically, the accuracy obtained with the combined model (75.2%) equals the accuracy obtained for the black South Africans (75.8%).

Table 6.7 – Confusion matrix showing patterns of overlap and misclassification among the groups for the combined training model employing both MMS traits and measurements.

		Classifies into:			%
		B	W	C	Correct
Group:	B	125	7	33	75.8
	W	3	152	10	92.1
	C	28	13	124	75.2
				Total:	81.0

The variable importance when comparing all variables simultaneously ranged from 1.0 to 33.4 (Table 6.8; Figure 6.3). The highest ranked variables include the inferior nasal margin (INA), nasal bone contour (NBC), nasal aperture shape (NAS), biasterionic breadth (ASB), nasal breadth (NLB) and the inter-orbital breadth measurement (DKB). While the three most important variables were MMS traits, the least important variables were also MMS traits, with nasal overgrowth (NO), post-bregmatic depression (PBD), and transverse palatine suture (TPS) contributing little information to the model.

Table 6.8 – RFM variable importance for the combined MMS traits and measurements.

Variable	Variable importance	Variable	Variable importance
INA	33.4	FRC	7.0
NBC	29.1	EKB	6.9
NAS	22.0	GOL	6.8
ASB	18.8	BBH	6.4
NLB	14.0	FOB	6.0
DKB	13.5	OBH	4.9
ANS	11.9	FOL	4.7
MDH	10.2	OCC	4.7
PAC	10.2	SPS	4.7
IOB	9.9	OBB	4.6
AUB	9.7	NBS	3.5
NLH	9.7	MT	3.3
NAW	8.3	PZT	3.1
ZS	8.1	NFS	2.8
UFBR	7.7	OS	2.8
BNL	7.6	TPS	2.3
XCB	7.6	PBD	1.2
ZYB	7.6	NO	1.0
WFB	7.2		

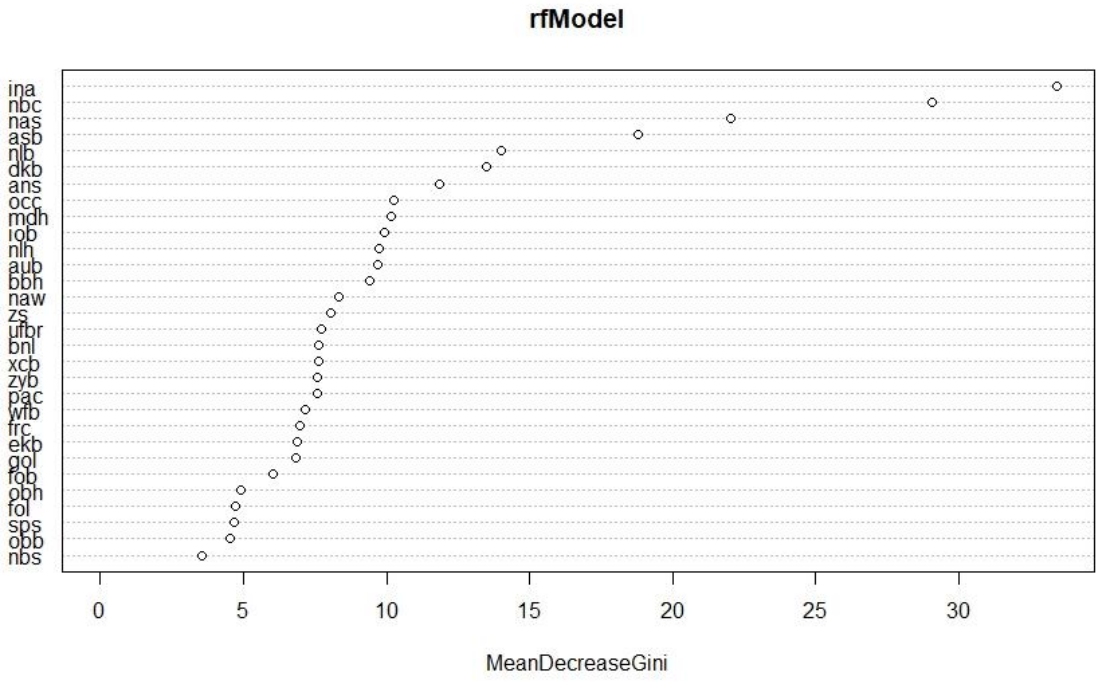


Figure 6.3 – Variable importance for the multivariate combined model employing all MMS traits and measurements.

To decrease the number of variables used in the model, all variables with an importance value below five were removed. Using 25 variables, the modified combined model yielded training and testing accuracies of 79.6% and 80.6%, respectively, with a kappa value of 70.9%. Table 6.9 provides a summary of all classification accuracies obtained with the multivariate models. It was consistently shown that removing variables decreased the accuracy of the models. Even though variables that did not demonstrate significant differences among the groups or that have low variable importance are included, every single variable contributes information to the model that facilitates better classification.

Table 6.9 – Comparison of the performance (%) of the MMS, measurement, and combined models to estimate ancestry.

Model	Training	Testing	Kappa
All MMS	78.8	81.8	72.7
All measurement	72.3	75.8	63.6
All combined	81.0	84.2	76.4
Modified MMS	73.6	75.25	62.9
Modified measurement	70.3	72.7	59.1
Modified combined	79.6	80.6	70.9

CHAPTER 7: DISCUSSION

Now more than ever, methods exploring ancestry need to be re-evaluated to ensure that valid methodology is employed, and that population variation is investigated and described in a scientifically meaningful way. As recommended by the Academy Standards Board (ASB) of the American Academy of Forensic Sciences, the estimation of ancestry should be based on peer-reviewed, published and validated methods that make use of appropriate reference samples. The combination of metric and non-metric data provides a detailed description of South African cranial variation which will assist practitioners to effectively quantify cranial size and shape to better understand and interpret results when using the cranium in skeletal analyses. The current study externally validates the MMS traits as a potential tool to estimate ancestry in South African anthropological analyses by providing population-specific data combined with robust quantitative analyses yielding high accuracies. However, certain aspects and limitations should be considered to further refine the method.

7.1 Methodology, observer agreement, and data variation

Non-metric methods are known to be subjective in nature and may produce variable results prone to bias (Hefner 2009; Hartley et al., 2022). This study attempted to explore some sources of trait score variation when scoring the MMS traits on the cranium for ancestry. Previous studies have referred to the implications of different statistics, observer experience and training, and population differences on consistently scoring morphological variation from the skeleton (Klales and Kenyhercz, 2015; Kamnikar et al., 2018; Klales et al., 2021).

Firstly, each statistical method used for reliability comparison have limitations which can create contradictory results. Klales and colleagues (2021) discuss the difficulties in comparing results from reliability studies of sex indicators as authors vary in the chosen statistics for observer agreement, ranging from Cohen's kappa with different weights, to intraclass correlations. While published studies typically make use of Cohen's kappa to assess the reliability of the MMS traits, a lack of consensus exists on the differential weighting of traits. The results of the current study demonstrated disparate rates of agreement depending on the weight assigned to the traits. Practitioners need to be cognisant of the fact that the MMS traits, while similar to the Walker (2008) or Klales et al. (2012) scoring systems, are not quite the same as these systems because all of the traits do not share the same data structure or number of trait expressions. No uniformly suitable Kappa weight can be applied to all the traits. The

variation in data structure needs to be considered when selecting tests for comparison and when interpreting results among multiple studies. The quadratic weighted kappa provides a realistic measure of agreement for the traits that are ordinally ranked with a logical order. Intuitively, as the ordinal scores are quasi-continuous with overlapping boundaries, being within one score should not be penalised as harshly as being two or more scores out. Among the most frequently cited literature, Hefner (2009), L'Abbé et al. (2011) and Klales and Kenyhercz (2015) did not specify whether any weights were assigned to the traits. Thus, unweighted kappa was likely employed, in which case the agreement for the ordinal traits may be underestimated. Conversely, Maier (2017) and Kamnikar et al. (2018) made use of quadratic-weighted kappa for all the traits. While it is suitable for the ordinal ranked traits, a quadratic weight is overly permissive for the nominal traits. Essentially, with the nominal scores there is no trait overlap, so whether an observer is within one score or not is irrelevant, as disagreement indicates misidentification of a particular shape or variant rather than misjudging the size of a skeletal feature. Applying a quadratic weight to all traits can result in underestimation of error for the nominal traits. Overall, the results indicate that different weights can influence the apparent reliability of a method, and practitioners should apply weights and tests that are suitable to the data type being analysed for each trait.

Additional limitations that have been associated with Cohen's kappa are issues of prevalence and bias, often referred to as the paradoxes of the kappa statistic. Prevalence occurs in cases where one trait state is much more prevalent than others, making it difficult to detect true agreement beyond chance (Byrt et al., 1993; Flight and Julious, 2015). Notably, prevalence might be a consequence of sampling. Using nasal overgrowth as an example: the trait was quite rare in the overall sample, but overgrowth was observed among less than 3% of black South Africans and was not observed at all among coloured South Africans. If a sample selected for reliability testing only included a group in which the trait is absent or exceedingly uncommon, the agreement score might not accurately reflect an observer's ability to score the trait correctly when present. On the other hand, bias refers to the frequency at which observers select a particular category (Byrt et al., 1993; Flight and Julious, 2015), and has to do with how an observer might interpret the descriptions and reference drawings. While prevalence can be mitigated with methodical sample selection, bias is difficult to control. Bias may result in failure to ascertain a coefficient with Cohen's kappa, as was noted with the first round of inter-observer scores prior to the group discussion. The calculation of additional prevalence and bias indices have been suggested to detect the paradoxes in a dataset; however, the indices are only

applicable to nominal data (Byrt et al., 1993). And while it is easy to calculate the indices for binary traits (i.e., scored as present or absent), the calculations become far more complex when a trait has more than two states. Alternatively, a prevalence and bias adjusted kappa (PABAK) has been proposed (Byrt et al., 1993), but again can only be applied to nominal data (Flight and Julious, 2015). The combination of several different data types into one method renders the analysis and comparison of traits challenging. No uniform approach is universally applicable, and practitioners need to demonstrate heightened awareness of strategies to circumvent potential issues pertaining to repeatability testing, especially when assessing non-metric traits. This includes the selection of sufficiently large samples; inclusion of the widest possible array of traits; selection of appropriate measures of repeatability that considers the characteristics of the data and number of states assigned to each trait; and providing sufficient detail to facilitate comparability of results across different studies.

Despite differences in the quantification of trait repeatability, results from the current study were compared to previously published research. The intra-observer agreement is equivalent to rates from other published studies (see Appendix VI for comparisons). Overall, three traits demonstrated moderate repeatability, which is the lowest rate of agreement recorded for the intra-observer analysis; the traits included inferior nasal margin (INA), nasal overgrowth (NO), and nasal bone shape (NBS). The inferior nasal margin is one of the traits with the greatest number of categories (with states from 1 to 5), where the trait expressions gauge whether the floor of the nasal aperture is smooth or sloping as it transitions to the maxilla, or whether the aperture is demarcated by a ridge a bone (Plemons and Hefner, 2016; Hefner and Linde, 2018). The change of the slope from one score to the next is gradual, and quite difficult to discern from photographs (Merchant, 2023). While studies involving the developer of the method observed substantial to almost perfect agreement (Hefner, 2009; Kamnikar et al., 2018), limited independent studies exist that report intra-observer rates. Both South African studies – the current study and that by L'Abbé and colleagues (2011) – demonstrated moderate agreement for the inferior nasal margin (INA). The lower rate of agreement compared to Hefner (2009) might have arisen due to misinterpretation of the description and images; however, it may also be the result of differences in trait expression attributable to population variation (Kamnikar et al., 2018). The expression of the inferior nasal margin (INA) in South Africans should be further explored given its consistent identification as the most discriminatory variable in the classification models; the repeatability of the inferior nasal margin (INA) is of paramount significance.

Both the nasal overgrowth (NO) and nasal bone shape (NBS) may be less repeatable because of subtle trait variations. The nasal overgrowth (NO) assesses projection of the nasal bones past the maxilla and is scored as either present or absent. Typically, the reliability of a trait is expected to be high if a variable only has two states which are sharply differentiated (McHugh, 2012). The disagreement in the evaluation of nasal overgrowth (NO) is likely related to the specific criteria concerning the extent to which the nasal bones must project beyond the maxilla to be scored as present. For the current study, when only a small portion of nasal bone projection was observed, the trait was scored as absent. Merchant (2023) raises similar concerns regarding the description of nasal overgrowth and how certain variations, such as separation of the nasal bones from the maxilla, correspond to the description of true nasal overgrowth. Finally, nasal bone shape evaluates the relationship between the degree of "pinching" and "bulging" that is exhibited by the nasal bones. In the current sample, the comparative drawings did not always align with the specimens; in numerous instances, the crania did not seem to match the extent to which the nasal bones "bulge" in the image. This could be indicative of variation among global populations. Additionally, the shape of the nasal bones was frequently noted to be asymmetrical. While Hefner and Linde (2018) contend that, in the case of asymmetry, the largest or most pronounced expression should be employed for numerous traits, this guideline is not specified for the nasal bone shape (Merchant, 2023). Contrary to the inferior nasal margin (INA), the nasal bone shape (NBS) and nasal overgrowth (NO) were not identified as highly discriminatory variables (although both traits do contribute information to the classification models). Nevertheless, the observed disagreements in the current study provide insights into the methodology for scoring these traits.

The level of inter-observer agreement is notably lower when compared to the intra-observer agreement, with several traits demonstrating poor repeatability. Disparity between inter- and intra-observer agreement should be ideally minimal, as notable discrepancies may suggest that while the intra-observer demonstrates consistency, it does not necessarily reflect the reliability of the method in terms of scoring the traits accurately. However, the trend of much greater intra-observer consistency when evaluating the MMS traits is continuously observed throughout the literature (Hefner, 2009; L'Abbé et al., 2011), and is likely the result of experience and familiarity with the method. The current study conducted a comparison based on levels of general experience, which revealed that less experienced individuals exhibited fewer traits with higher levels of agreement. This observation is not necessarily reflected in the mean kappa values calculated per observer, as even observers with more extensive general

experience displayed comparably low repeatability. Collectively, none of the traits consistently exhibited poor repeatability across all observers. Instead, each observer presented with different traits that achieved the highest and lowest levels of repeatability, respectively. The variability among observers suggests that in the absence of additional guidance or shared knowledge, each observer resorts to individualised scoring approaches for different traits. This assumption is substantiated by potential instances of scoring bias, as is evidenced by the numerous cases where it was not possible to calculate a kappa value for certain traits. Thus, general experience with skeletal material does not directly translate to competency in using the traits.

Multiple authors have underscored the need for method-specific training to use the MMS traits (Klales and Kenyhercz, 2015; Kamnikar et al., 2018). While certainly not equivalent to continuous comprehensive training, a discussion session was conducted with all the observers to assess whether even a modest degree of familiarity and additional instruction on the scoring procedure could improve reliability results. The group discussion appeared to have a positive impact, as several traits demonstrated increased repeatability compared to the previous scores. Although, this improvement is once again not reflected in the mean kappa values for each observer. The discussion culminated in mixed results. Two observers (C and E), demonstrated greater repeatability following the discussion, resulting in mean kappa values that progressed from fair to moderate. Interestingly, these individuals have mixed levels of general experience, suggesting that general experience does not contribute significantly to a more consistent assessment of the traits. In contrast, the other two observers (B and D) demonstrated decreased repeatability in comparison to the first round of scores. This result initially seemed unexpected, as Observer B is the only other observer with prior experience with the traits, having previously published on the subject. However, closer examination revealed that their experience pertained to the traits as described in the original publication (Hefner, 2009). Since the initial publication, several modifications have been introduced, including the incorporation of additional traits, adjustment to the trait scales, and the implementation of the *MMS* user interface, all of which may influence the observer agreement. Furthermore, very limited research has aimed to quantify the repeatability of traits over an extended period of time, which may also contribute to lower agreement despite experience with the traits (Kamnikar et al., 2018).

During the second round of scoring, no ‘NaN’ values were encountered (i.e., instances where kappa values could not be calculated), indicating that the potential scoring bias may have been somewhat mitigated after discussing the traits. Method discussions among practitioners

are essential, as it can enhance comprehension of standard procedures, facilitate terminology-related deliberations, and shed light on the implication of language, translation and personal interpretation of terms on quantifying non-metric variation (Wilczak et al., 2017). Furthermore, such discussions provide insight into the various approaches taken by practitioners to resolve issues related to trait assessment which may lead to error or bias, especially in cases where there are no established guidelines to address certain trait variations.

In the current study, several personal approaches to scoring the traits became apparent. Certain observers (e.g., observer A – the principal investigator) placed a significant emphasis on tactile examination to assess the size of some traits, such as the anterior nasal spine (ANS), nasal aperture width (NAW), and posterior zygomatic tubercle (PZT). The repeatability of the above traits improved when the other observers adopted this approach. The use of different tools to visually assess certain traits also varied among the observers. In an attempt to improve trait repeatability, the most recent guidelines recommend the use of a contour gauge to better visualize the nasal bone contour, and a clear ruler is recommended to examine the size of the malar tubercle and posterior zygomatic tubercle (Plemons and Hefner, 2016; Hefner and Linde, 2018; Kamnikar et al., 2018). One observer commented on using the ruler to also assess the nasal aperture width and inter-orbital breadth, which essentially converts the trait to a measurement, and would be more accurately measured with a caliper. Merchant (2023) addresses ambiguity regarding the location and placement of the ruler to assess the posterior zygomatic (PZT) and malar tubercles (MT), ultimately highlighting a lack of consensus among their cohort of observers. While the exact placement of the ruler did not form part of the collective discussion in the current study, this omission is likely attributed to the observers using it infrequently. Throughout the training period, Observer A attempted scoring both with and without the ruler and observed greater consistency when it was not used. Since Observer A was responsible for collecting all the data used throughout the current study, the ruler was not used as part of the scoring procedure. Similarly, the contour gauge received limited preference to score nasal bone contour (NBC), and it was not used during data collection. In the group discussion, several observers noted that scoring the trait with the contour gauge consistently yielded the same score (despite the nasal region itself looking different), leading to repeatability poorer than chance and introducing bias in the scores. Following additional instructions provided by Observer A, subsequent attempts to score nasal bone contour (NBC) without the contour gauge demonstrated improved repeatability; however, the kappa value remained quite low. The results support findings in the literature calling for training prior to

using the traits in research or skeletal analyses (Klales and Kenyhercz, 2015; Kamnikar et al., 2018).

Although the importance of training cannot be overstated, Wilczak and colleagues (2017) raise concerns regarding the potential implications of "second-hand" and self-training in scoring. Typically, developers of new methods offer training at workshops or through collaborative projects. However, as methods become more established and widely applied, the availability of training opportunities diminishes. As such, practitioners often need to rely on published descriptions and photographs, or training provided by independent individuals with some experience in the method. Although observers without direct training from developers can still produce consistent results (especially for intra-observer agreement), the possibility exists that discrepancies may arise compared to the developers or other experts in the field (Wilczak et al., 2017). Such discrepancies can lead to variations in trait frequencies between studies, and ultimately decreased classification accuracy (Lewis and Garvin, 2016; Klales et al., 2020). Furthermore, discrepancies can become standard practice as it is passed down from one generation to the next through educational pedagogy (Klales, 2021). Additional research needs to evaluate the precision and reliability of scoring the MMS traits, especially for the sake of data sharing and the collation of a global database. Ultimately, the observer agreement achieved in this research is in line with previous studies evaluating the MMS traits, which have considered the repeatability satisfactory to justify its use in practice. Nevertheless, it is important to acknowledge that there is no established threshold for what constitutes a kappa value that is acceptable. Further deliberation is necessary to establish criteria for adequate levels of validity and reliability by which a method can be assessed for its applicability in forensic casework (Klales, 2021). In order for the method to be a viable option to conduct ancestry estimation in South Africa, the forensic anthropology community responsible for assessing skeletal remains should be subjected to rigorous training in scoring the MMS traits prior to the method being used in analyses.

The collection of measurements is considered much more objective and repeatable than non-metric methods as it makes use of standard landmarks with clear definitions and calibrated tools. A multitude of studies have examined the repeatability of measurements on the cranium, postcranium, and across various digital modalities (Adams and Byrd, 2002; Franklin et al., 2013; Stull et al., 2014b; Smith and Boaks, 2017; Langley et al., 2018; Liebenberg and Krüger, 2020). While measurements are more objective, they are not completely free from error (Hartley et al., 2022. The most prevalent source of measurement error stems from ambiguity

surrounding the precise identification of landmark locations. Type I landmarks, positioned at the intersection of structures such as sutures, typically yield the least measurement error (Smith and Boaks, 2017). The prevalence of type I landmarks on the cranium may contribute to the perception of the cranium as having a higher level of repeatability compared to the postcranial skeleton. In the present study, measurement errors were found to be consistently low, with all measurements falling within the conventional margin of error of ±2mm, as accepted in the field of anthropology (Stull et al., 2014b; Smith and Boaks, 2017). However, certain measurements, such as inter-orbital breadth and mastoid height, exhibited greater measurement variation. Additionally, greater error was also recorded for the maxilla-alveolar length and breadth, as well as the length of the foramen magnum. The inter-orbital breadth (measured from dacryon to dacryon) has previously been identified as particularly susceptible to error in multiple studies (Franklin et al., 2013; Stull et al., 2014b; Smith and Boaks, 2017). Even though dacryon is classed as a type I landmark, Smith and Boaks (2017) reported poor agreement on the exact location of this landmark. In the case of mastoid height, the measurement error may arise from personal idiosyncrasies during data collection, such as variations in caliper orientation (Smith and Boaks, 2017). Similar to the MMS traits, no universally accepted threshold exists as to what constitutes an acceptable TEM value for measurements. Nevertheless, the measurements in this study achieved agreement levels consistent with those found in other studies and were thus deemed satisfactory.

7.2 Population variation and classification

Skeletal variation attributable to ancestry has been shown to be highly variable, not only among different populations across the globe, but also within populations and population groups (Ousley et al., 2009). The variation observed among the three South African population groups has been discussed in terms of their population histories, which were significantly influenced by migration, colonisation, and institutionalised racism (Stull et al., 2016; Krüger et al., 2018). The current study revealed substantial group overlap in the crania of modern black, white and coloured South Africans, which aligns with findings in previous studies (L'Abbé et al., 2011; L'Abbé et al., 2013; Stull et al., 2014a; Liebenberg et al., 2015b). Both metric and MMS data reveal consistent patterns of misclassification, where coloured South Africans misclassify nearly equal with both black and white South Africans. In contrast, black and white South Africans rarely misclassify as one another. These findings align with the assertion by Hefner and colleagues (2014) that craniometric and MMS data yield similar insights into the

relationships between and among populations. Coloured South Africans are typically reported to exhibit the lowest classification accuracy when compared to black and white South Africans, particularly in cranial analyses. This increased misclassification has been linked to their complex genetic composition (Adhikari, 2005), and the intermediacy in terms of cranial morphology relative to the other groups. Coloured South Africans have been shown to share similarities with white South Africans in cranial size but display greater similarities with black South Africans in cranial shape (Stull et al., 2016; Krüger et al., 2018). Despite the substantial overlap, various traits and measurements demonstrate significant differences across all three groups, implying the potential for group differentiation when employed in multivariate analyses. This was subsequently validated with the mixed classification model, which showed greater accuracy when classifying coloured South Africans compared to either the MMS or metric models on their own. Thus, the combination of size and shape variables in the mixed model proves to be more effective in distinguishing coloured South Africans from the other population groups, resulting in reduced misclassification and improved predictive performance. Future research needs to further explore the craniofacial variation and overlap of the South African coloured group; given their complexity, alternative subdivisions within the group may also be considered to better capture the great amount of variation that is observed among coloured individuals.

The midfacial region of the cranium is frequently cited as the most discriminative area for the estimation of ancestry (Brues, 1990; McDowell et al., 2012, 2015; Liebenberg et al., 2015b). The craniofacial complex is a modulated structure in which adjacent bones interact under the influence of developmental, genetic, and functional factors to give to a highly integrated phenotype (Bastir et al., 2006; Martínez-Abadías et al., 2012). In essence, changes in one skeletal element are mirrored by corresponding changes in adjacent areas (Bastir, 2008). In the context of the facial skeleton, the nasal bones, maxillae, and zygomas share a close relationship. Their proximity is expected to result in stronger correlations among the morphological features and dimensions, thereby leading to covariance within and among populations (Martínez-Abadías et al., 2012; Mitteroecker et al., 2012). Many variables examined in the study are concentrated in the facial region, which demonstrated particular patterns of variation. For example, crania with a larger nasal width also exhibited a wider inter-orbital breadth, and a bowed nasal aperture shape. Similarly, crania with a teardrop-shaped nasal aperture were likely to exhibit a sharper inferior nasal margin. The results also identified inverse relationships; crania with a bowed nasal aperture are less likely to have a large nasal

spine or sharp nasal margin, while crania with wide inter-orbital breadths and/or bowed nasal apertures are less likely to present with nasal overgrowth. Despite the apparent covariation, the MMS traits did not display strong correlations with one another or with the cranial measurements. The lack of strong correlations is somewhat unexpected, considering the proximity of the traits and the fact that certain variables quantify the same feature, such as the score and measurement for interorbital breadth (IOB and DKB), as well as the score and measurement for nasal width/breadth (NAW and NLB). The moderate strength of the correlations may indicate a limited relationship among the features; however, measurements quantifying size are recognised to be more proficient in capturing such relationships than non-metric traits (Mitteroecker et al., 2012). When comparing the measurements with one another, a higher number of variables displayed stronger correlations. More appropriate analyses, such as maximum likelihood methods, should be used to provide more comprehensive insights on the variable relationships. While the covariance and heritability of cranial measurements has been extensively examined, the heritability of non-metric traits, particularly within the context of the MMS method, remains incompletely understood to date (Corrucini, 1974; Relethford, 1994; Carson, 2006; Martínez-Abadías et al., 2009). Ross and Pilloud (2021) contend that a more biological perspective should be applied to the evaluation of MMS traits, wherein heritability and evolutionary significance require further exploration.

The findings of the current study confirm the premise that the midface, and specifically the nasal region, plays a pivotal role in ancestry estimation. The variables not only demonstrated significant differences, with many showing marked differences among all three groups assessed, but also proved to be beneficial within the classification models. For group classification, both univariate and multivariate analyses were conducted. While researchers widely acknowledged that multivariate analyses outperform single variables (Ousley and Jantz, 2012), the performance of individual variables is essential to understand their ability to estimate ancestry when limited skeletal material is available, like with fragmentary crania. The majority of the variables displayed relatively moderate accuracies for the univariate analyses, with the univariate MMS models exhibiting slightly higher accuracy compared to univariate measurements. The nasal bone contour (NBC) and inferior nasal margin (INA) demonstrated high accuracies (69% and 66%, respectively), which is notable considering the models represented single traits distinguishing among three groups. In comparison, the highest accuracies observed with measurements were noted for nasal breadth (NLB) and interorbital breadth (DKB) (54% and 52%, respectively). This suggests that the MMS traits are more

effective at capturing variation among groups, whereas the cranial measurements exhibit greater similarity and overlap among the groups, which presents difficulty in defining sufficient boundaries for group classification.

The multivariate analyses further substantiate this assumption, as the MMS model outperformed the measurement model. The measurement model achieved accuracies comparable to previously published error rates for the classification of the South African groups using standard craniometrics with discriminant analysis (L'Abbé et al., 2013). While the predictive performance is better than chance, the measurements allow for a notable margin of error. This is likely because much of the variation associated with the cranium is not quantified effectively when applying linear distances to measure a round object. The mixed model achieved the greatest results (81% to 84% correct classification), which align with established postcraniometric standards (Liebenberg et al., 2015a) and morphometric data employing cranial features for the same South African population groups (Stull et al., 2014a). Variable importance analysis indicated that the mixed model heavily relied on variables that assess facial shape (such as inferior nasal margin - INA, nasal bone contour - NBC, and nasal aperture shape - NAS) in conjunction with measurements that quantify both facial size (nasal breadth – NLB, and interorbital breadth - DKB) and cranial vault size (biasterionic breadth – ASB, and parietal chord - PAC). Thus, the two datasets capture the variation of the cranium differently and a comprehensive assessment of both size and shape is required to achieve the best results for cranial ancestry estimation (Stull et al., 2014a). Many authors have documented the superior results attainable through mixed models (e.g., Hefner et al., 2014; Maier, 2019; Klales, 2020). Maier (2019) highlights the benefit of gathering more information through simultaneous analysis of multiple datasets, emphasising that mixed models can offer improved assessment of variation in complex groups, such as Hispanic individuals in the United States (and by extension, the highly heterogeneous coloured South Africans in the current study). An additional advantage of mixed models is that skeletal variation from multiple methods can be integrated into a single ancestry estimate. The interpretation and collation of results from multiple methods has been shown to vary substantially among forensic practitioners, often without any empirical basis (Garvin and Passalacqua, 2012; Klales, 2021). This can ultimately lead to disparate results and introduce bias in forensic reports (Hartley et al., 2022). Currently, practitioners offer no consensus on how to combine results from different methods into final estimates for forensic reports (Klales, 2021).

This study supports previous research in stating the great potential of RFM as a classification method (Hefner et al., 2014; Navega et al., 2015; Maier, 2019; Klales, 2020). As RFM is non-parametric, the method does not rely on statistical assumptions like normality, which are rarely met in real-world data. The method is capable of combining different types of data and includes internal validation functionality which eliminates the need for additional independent samples to test the model validity. Finally, RFM is not prone to overfitting and the curse of dimensionality, which is a well-known issue encountered with discriminant analysis (Ousley and Jantz, 2012). The inclusion of a greater number of measurements is recognised to allow more differences to be detected among groups. However, a decrease in classification accuracy will often be noted as more variables are added (Ousley, 2016). Essentially, redundant and highly correlated variables introduce statistical "noise", which adversely affects the predictive performance of a model. The solution to this problem is to reduce the number of variables so that only the most discriminatory variables are retained. For example, with linear discriminant analysis, stepwise selection is employed as a variable reduction technique (Ousley and Jantz, 2012; Ousley, 2016). RFMs are capable of handling large numbers of variables, and it has been recommended that as many variables as possible be included and the model be allowed to run with them (Hefner and Ousley, 2014; Navega et al., 2015). Navega and colleagues (2015) specifically caution against removing variables, even if they exhibit low measures of variable importance. Variable importance reflects the contribution of a specific trait or measurement to the overall ensemble of trees used in the model. In the current study, the ensemble consisted of 2500 trees. However, each individual tree employs a random subset of variables at each split. Consequently, the overall contribution to the model may appear small, but the variable importance does not necessarily reflect how discriminative a variable can be for certain individual trees within the ensemble (Navega et al., 2015). Indeed, the current study demonstrated that the removal of even a single variable led to decreased accuracy. A notable strength of RFM is its efficiency in capturing interactions between variables as the model tests different combinations at each split, which makes it a highly effective classification tool with strong generalization capabilities (Navega et al., 2015).

7.3 Practical application and future recommendations

This study confirmed the potential of MMS traits in exploring ancestry and the positive results that the method can yield, particularly when used alongside craniometrics. However, the findings also identified several areas that warrant further exploration. These include

additional covarying factors that may influence cranial morphology and how the traits are expressed, such as sexual dimorphism, age, edentulism, and asymmetry.

Previous research has examined metric sex differences in the crania of South Africans (e.g., Steyn and İşcan, 1998; Franklin et al., 2005; Dayal et al., 2008; Small et al., 2018). The results reveal significant differences between males and females, resulting in high classification accuracies. However, these studies evaluated population groups in isolation (i.e., only looking at either black or white South Africans) without simultaneously comparing multiple different population groups and sexes to comprehensively assess the interaction of sex and ancestry on cranial morphology. In a morphoscopic study, Krüger et al. (2015) identified significant differences between black and white South Africans using the Walker (2008) traits, and thus supported the need for population-specific standards to estimate sex. Furthermore, correlations between the Walker (2008) traits and cranial measurements revealed strong relationships, suggesting sex plays a role in both the size and shape of the cranium (Krüger et al., 2015). L'Abbé et al. (2013) simultaneously considered sex and ancestry when attempting to estimate ancestry with cranial measurements and observed that the cranium frequently misclassified according to sex. Thus, the close variation between sex and ancestry can affect the positive predictive performance of the cranium in correctly assigning sex and ancestry. Concerning the MMS traits, Hefner (2009) reported no significant sex differences, suggesting that the sexes be pooled for further analyses. However, sex has previously been shown to have a significant impact on inter-orbital breadth (IOB) in a South African population (L'Abbé et al., 2011). Similarly, the current study observed significant sex differences for several traits, including the inferior nasal margin (INA), inter-orbital breadth (IOB), malar tubercle (MT), nasal aperture width (NAW), posterior zygomatic tubercle (PZT), and supra-nasal suture (SPS). The South African population has demonstrated varying levels of sexual dimorphism compared to North Americans (Caple and Stephan, 2017), which could account for the significant sex differences observed with the MMS traits. Conducting sex-specific analyses can mitigate the impact of sexual dimorphism, allowing classification models to focus solely on assessing differences related to ancestry. Prior knowledge of sex can enhance classification accuracy by reducing group overlap, thereby facilitating more effective group separation (Liebenberg et al., 2019). Further studies are warranted to explore whether sex-specific analyses would yield improved results for the MMS traits.

Similarly, the effects of age on craniofacial morphology need to be considered. Ontogeny and cranial growth in subadults have been extensively documented, with cranial stasis

(cessation of growth) reported around 17 years of age (Ross and Williams, 2010). The changes in the craniofacial complex throughout the adult lifespan remain less thoroughly understood. Existing literature suggests that cranial remodelling persists throughout adulthood, revealing differences in both cranial size and shape (Akgül and Toygar, 2002; Albert et al., 2007; Patterson et al., 2007). Ross and Williams (2010) describe a general expansion in facial breadth, with certain craniometric landmarks around the bony orbit – such as frontomalare temporale, dacryon and ectoconchion – shifting laterally as age advances. This lateral movement results in enlarged orbits (Williams and Slice, 2010). Additional observations indicate an increase in cranial circumference, elongation and widening of the face, and lengthening of the mandible with an increased mandibular angle (Albert et al., 2007; Ross and Williams, 2010; Williams and Slice, 2010). Apart from facial changes, an enlargement of the cranial base has been noted, involving a downward shift of the external occipital protuberance, lambda, and the mastoid processes, while asterion has been noted to shift posteriorly (Ross and Williams, 2010; Small et al., 2016). Notably, these are all prominent landmarks used in standard craniometric methods. The most conspicuous facial changes are observed in the dento-alveolar region, primarily influenced by antemortem tooth loss. Advancing age contributes to bone porosity due to decreased osteoblastic activity, which increases the risk of tooth loss (Small et al., 2016). The overall remodelling of alveolar bone due to edentulism produces a concave facial appearance along with an increased facial height, regression of the dento-alveolar region, and retrusion of the maxilla (Albert et al., 2007; Nikita, 2014). Edentulism may also lead to asymmetry in the cranium, particularly in instances of unilateral tooth loss and asymmetrical mastication (Dinkele, 2018).

The majority of previous studies employed morphometric techniques to investigate age-related changes, but few authors have assessed the implications of age and edentulism on morphoscopic variation. Regarding sexual dimorphism, older females have been reported to exhibit more robust cranial features (Walker, 1995). Although age shows a significant correlation with cranial morphoscopic traits, further research indicated that age did not significantly affect sex estimation using these traits (Klales, 2021). A limited selection of the MMS traits for ancestry have been explored for age variation. The anterior nasal spine has been found to exhibit a significant relationship with age, where older individuals tend to display longer, more pronounced spines (L'Abbé et al., 2011; Dinkele, 2018). However, the relationship between the remainder of the traits and age remains speculative (Caple and Stephan, 2017), and more importantly, whether or not this association with age will impact the

predictive performance of the traits in ancestry estimation is unknown. Age and edentulism posed concerns in the current study, given the disparity in the mean age of white South Africans compared to the other groups. Despite the mean age difference, both the white and coloured South African groups exhibited substantial amounts of edentulism. Indeed, the presence of edentulism not only impacts cranial features, as discussed earlier, but it also restricts the sample size by preventing the collection of variables related to the palate, owing to significant alveolar resorption. Additional classification models should be created to assess if age-specific models would yield greater classification accuracies. Furthermore, the number and patterns of missing teeth should be assessed in the current sample to explore if there is a relationship between age, edentulism and the cranial variables (especially the MMS traits), given the paucity of information on this subject.

Finally, additional strategies need to be explored to improve the observer repeatability of the traits. Method-specific training remains a top priority for mastering the MMS method (Klales and Kenyhercz, 2015). However, the specific components that should be included in the training process has yet to be clearly defined. Direct training from the method developer would be the optimal way forward; however, this is simply not feasible. Even if continuous training opportunities were offered, international conferences and workshops are expensive and not all practitioners and students may have the financial resources to attend. The availability of online resources can be useful in promoting consistency in scoring traits globally and in addressing discrepancies in a standardised manner. In a comprehensive study examining the qualitative assessment of pathological lesions on bone, Wilczak and colleagues (2017) recommended several steps to potentially improve confidence and validity in scoring methods. These steps included on-going discussions on terminology and method refinement; making exemplar cases and case studies of trait variations available; and incorporating 3D models into the training and scoring procedure (Wilczak et al., 2017). While Hefner and Linde (2018) have published a photographic atlas showcasing the different MMS traits, challenges related to scoring the traits from 2D photographs have been discussed (Merchcant, 2023). Other authors have also highlighted issues with scoring from 2D photographs, such as poor lighting, sub-optimal angles in photographs, and the skeletal feature not being entirely in the plane of focus (Stephan and Caple, 2017; Wilczak et al., 2017; Craik and Collings, 2022; Merchant, 2023). This makes scoring particularly challenging for individuals that are relying on photographs for training purposes.

The use of 3D models is a potential solution for the challenges associated with 2D photographs. With technological advancements, forensic anthropologists are increasingly integrating 3D virtual reconstructions of bones into their research, educational practices, and forensic evidence reconstruction for legal testimony (Carew et al., 2019; Craik and Collings, 2022). In virtual 3D models, users can manipulate the image by rotating the bone or zooming in on specific features. The viewing software also allows for the application of different settings, such as varied lighting or textures, to enhance specific traits or features. Incorporating 3D models into teaching has proven effective in improving comprehension and consistency among students in identifying skeletal features (Craik and Collings, 2022). While 3D models offer several advantages, they are not considered a complete replacement for physical bones, particularly for traits that require palpation. Kuzminsky et al. (2020) noted that although virtual models were suitable for experienced individuals in analysing sex from the cranium, they prove less reliable for individuals with less experience. Thus, a combined approach utilising real bones, 3D printed bones or casts alongside virtual models may yield optimal training. The impact of using 3D models (both virtual and physical) as part of MMS trait training should be further explored.

CHAPTER 8: CONCLUSION

The current study is the first to conduct a comprehensive analysis of MMS variation and predictive performance in a modern South African population. Numerous exploratory analyses were conducted to show that despite substantial heterogeneity and overlap, sufficient cranial differences exist among black, white and coloured South Africans to be able to estimate ancestry using the MMS traits. Ultimately, the classification models demonstrated that MMS traits outperform craniometric techniques currently employed for ancestry estimation. This confirms that the variation in the craniofacial complex results from both size and shape differences, an aspect more effectively quantified with MMS traits compared to cranial measurements, which predominantly assess size. The findings validate the use of MMS traits as a tool to estimate ancestry in South Africa; thus, the method can be incorporated into South African forensic casework. The South African-specific database created in this study will be submitted for inclusion in the global MaMD databank. Once integrated, the *MMS* analytical tool associated with the MaMD can be employed to assess the ancestry of unknown South Africans.

Moreover, this study demonstrated that a mixed model, incorporating both MMS traits and measurements, achieved high accuracies surpassing those of each method when used on its own. This further underscores the effectiveness of combining numerous size and shape variables, especially in assessing complex, heterogeneous groups characterised by substantial within-group variation and overlap. Due to their size overlap with white South Africans, and shape overlap with black South Africans, coloured South Africans are often prone to misclassification, resulting in significantly lower classification accuracies. However, the mixed model successfully classified coloured South Africans with comparable accuracy to the other groups, a challenge that limited other methods and studies. While the combination of variables contributed significantly to the improved group separation, the choice of classification statistics also played a crucial role.

Random forest modelling (RFM) proved to be a valuable tool to analyse the mixed model, offering several advantages. It can be applied to any data type including mixed data with combinations of continuous, nominal, and ordinal variables. The models are robust against statistical assumptions (such as normality) and overfitting and can handle a large number of variables. Despite being computationally robust, RFM is overall user-friendly and allows the

inclusion of large amounts of variables without concerns about additional or manual variable reduction. The successful performance of the mixed model, paired with the advantages of RFM, indicates a need for a computer software program that will allow the case-by-case analysis of South African crania using RFM with both MMS traits and measurements, which is not possible with the *MMS* analytical tool.

While promising results were noted, some areas for improvement were also identified. Other factors that influence the morphology of the craniofacial complex, including sex, age and edentulism, should be explored to gain a better understanding of the sources of variation in the cranium and the implications on estimates of the biological profile. Finally, observer repeatability remains a concern for the method to be applied successfully. Method-specific training should be a requirement before the method is used. Additional tools such as 3D models should be assessed in an attempt to enhance the trait repeatability. Ultimately, method refinement and validation should be an on-going, globally collaborative effort.

REFERENCES

Adams BJ, Byrd JE. 2002. Interobserver variation of selected postcranial skeletal measurements. J Forensic Sci 47: 1193-1202.

Adhikari M. 2005. Contending approaches to coloured identity and the history of the coloured people of South Africa. Hist Compass 3:1-6.

Akgül AA, Toygar TU. 2002. Natural craniofacial changes in the third decade of life: A longitudinal study. Am J Orthod Dentofacial Orthop 122: 512 – 522.

Albert AM, Ricanek Jr. K, Patterson E. 2007. A review of the literature on the adult aging skull and face: Implications for forensic research and applications. Forensic Sci Int 172: 1 – 9.

Alblas A, Greyling LM, Geldenhuys EM. 2018. Composition of the Kirsten Collection at Stellenbosch University. S Afr J Sci 114:1-6.

Ali Z, Bhaskar SB. 2016. Basic statistical tools in research and data analysis. Indian J Anaesth 60: 662-669.

Allan A, Louw DA. 2001. Lawyers' perception of psychologists who do forensic work. S Afr J Psychol 31:12-20.

Bass WM. 1995. Human osteology: a laboratory and field method. Charles C. Thomas: Springfield, IL.

Bastir M, Rosas A, O'Higgins P. 2006. Craniofacial levels and the morphological maturation of the human skull. J Anat 209: 637-54.

Bastir M. 2008. A systems-model for the morphological analysis of integration and modularity in human craniofacial evolution. J Anthropol Sci 86: 37-58.

Berry AC, Berry RJ. 1967. Epigenetic variation in the human cranium. J Anat 101: 361.

Bethard JD, DiGangi EA. 2020. Moving beyond a lost cause: Forensic anthropology and ancestry estimates in the United States. J Forensic Sci 65: 1791.

Brace CL. 1964. On the race concept. Curr Anthropol 5: 313-320.

Breiman L. 2001. Random forests. Mach Learn 45: 5-32.

Brues AM. 1990. The once and future diagnosis of race. In: Gill GW and Rhine S (eds). Skeletal attribution of race: Methods for forensic anthropology. University of New Mexico, Albuquerque. p.1-7.

Byrt T, Bishop J, Carlin JB. 1993. Bias, prevalence and kappa. J Clin Epidemiol 46: 423-429.

Caple J, Stephan CN. 2017. Photo-realistic statistical skull morphotypes: new exemplars for ancestry and sex estimation in forensic anthropology. J Forensic Sci 62: 562-572.

Carew RM, Morgan RM, Rando C. 2019. A preliminary investigation into the accuracy of 3D modeling and 3D printing in forensic anthropology evidence reconstruction. J Forensic Sci 64: 342-352.

Carson EA. 2006. Maximum-likelihood variance components analysis of heritabilities of cranial nonmetric traits. Hum Biol 78: 383-402.

Caspari R. 2003. From types to populations: A century of race, physical anthropology, and the American Anthropological Association. Am Anthr 105: 65-76.

Caspari R. 2009. 1918: Three perspectives on race and human evolution. Am J Phys Anthropol 139:5-15.

Chan YH. 2003. Biostatistics 104: correlational analysis. Singapore Med J 44: 614-619.

Christensen AM, Crowder CM. 2009. Evidentiary standards for Forensic Anthropology. J Forensic Sci 54:1211-1216.

Christensen AM, Passalacqua NV, Bartelink EJ. 2013. Forensic anthropology: Current methods and practice. Elsevier Inc. p. 223 – 24.

Christensen AM, Smith MA, Gleiber DS, Cunningham DL, Wescott DJ. 2018. The use of X-ray computed tomography technologies in forensic anthropology. J Forensic Anthropol 1: 124 - 140.

Committee on Identifying the Needs of the Forensic Sciences Community, National Research council (NAS). 2009. Strengthening forensic science in the United States: A path forward.

Coon CS, Garn SM, Birdsell JB. Races: A study of the problems of race formation in man. Charles C. Thomas, Springfield, IL. p. 65-71.

Corruccini RS. 1974. An examination of the meaning of cranial discrete traits for human skeletal biological studies. Am J Phys Anthropol 40: 425-445.

Craik K, Collings AJ. 2022. A preliminary study into the impact of using three-dimensional models in forensic anthropology learning and teaching. Sci Justice 62: 814-821.

Curtis EA. Comiskey C, Dempsey O. 2016. Importance and use of correlational research. Nurse Res 23: 20-25.

Daubert v. Merrell Dow Pharmaceuticals, Inc, 509 US. 579. 1993.

Dayal MR, Spocter MA, Bidmos MA. 2008. An assessment of sex using the skull of black South Africans by discriminant function analysis. Homo 59: 209-221.

De Villiers H. 1968. The skull of the South African Negro: A biometrical and morphological study. Witwatersrand University Press, Johannesburg.

DiGangi EA, Bethard JD. 2021. Uncloaking a lost cause: Decolonizing ancestry estimation in the United States. Am J Phys Anthropol 175: 422-436.

Dinkele E. 2018. Ancestral variation in mid-craniofacial morphology in a South African sample [Masters dissertation]. University of Cape Town.

Dirkmaat DC, Cabo LL, Ousley SD, Symes SA. 2008. New perspectives in forensic anthropology. Am J Phys Anthropol 137:33-52.

Dirkmaat DC, Cabo LL. 2012. Forensic anthropology: Embracing the new paradigm. In: Dirkmaat DC (ed). A companion to Forensic Anthropology. Blackwell Publishing LTD. p.3-40.

Dubow S. 1995. Scientific racism in modern South Africa. Cambridge University Press.

Dubow S. 1996. Human origins, race typology and the other Raymond Dart. Afr Stud 55: 1-30.

Dunn RR, Spiros MC, Kamnikar KR, Plemons AM, Hefner JT. 2020. Ancestry estimation in forensic anthropology: A review. WIRES: Forensic Science 2: e1369.

Edgar HJH, Hunley KL. 2009. Race reconciled: How biological anthropologists view human variation. Am J Phys Anthropol 139:1-4.

Edgar H, Pilloud, M. 2021. A Reassessment of Assessing Race: " Ancestry" Estimation and Its Implications for Forensic Anthropology and Beyond. J Forensic Anthropol 4: 67-72.

Fabris F, Doherty A, Palmer D, de Magalhães JP, Freitas AA, Wren J. 2018. A new approach for interpreting random forest models and its application to the biology of aging. Bioinformatics 16: 1-8.

Ferrante L, Cameriere R. 2009. Statistical methods to assess the reliability of measurements in the procedures for forensic age estimation. Int J Legal Med 123:277-283.

Flight L, Julious SA. 2015. The disagreeable behaviour of the kappa statistic. Pharm Stat 14: 74-78.

Fox J. 2022. polycor: Polychoric and polyserial correlations. Retrieved from: https://CRAN.R-project.org/package=polycor.

Franklin D, Freedman L, Milne N. 2005. Sexual dimorphism and discriminant function sexing in indigenous South African crania. Homo *55:* 213-228.

Franklin D, Cardini A, Oxnard CE. 2010. A geometric morphometric approach to the quantification of population variation in sub-Saharan African crania. Am J Hum Biol 22:23-35.

Franklin D, Cardini A, Flavel A, Kuliukas A, Marks MK, Hart R, Oxnard C, O'Higgins P. 2013. Concordance of traditional osteometric and volume-rendered MSCT interlandmark cranial measurements. Int J Legal Med 127: 505-520.

Franklin D, Blau S. 2020. Physical and virtual sources of biological data in forensic anthropology: Considerations relative to practitioner and/or judicial requirements. In: Obertova Z, Stewart A, Cattaneo C (eds). Statistics and probability in forensic anthropology. Academic Press p 17-45.

Gamer M, Lemon J, Fellows I, Singh P. 2019. irr: Various coeffiecients of interrarter reliability and agreement. Retrieved from: https://CRAN.R-project.org/package=irr.

Garvin HM, Passalacqua NV. 2012. Current practices by forensic anthropologists in adult skeletal age estimation. J Forensic Sci 57: 427-433.

Garvin HM, Stock MK. 2016. The utility of advanced imaging in forensic anthropology. Acad Forensic Pathol 6: 499-516.

Gill GW. 1990. Introduction. In: Gill GW and Rhine S (eds). Skeletal attribution of race: Methods for forensic anthropology. University of New Mexico, Albuquerque. p.vii-xii.

Grivas CR, Komar DA. 2008. *Kumho, Daubert* and the nature of scientific inquiry: Implications for Forensic Anthropology. J Forensic Sci 53:771-776.

Hartley S, Winburn AP, Dror, IE. 2022. Metric forensic anthropology decisions: Reliability and biasability of sectioning-point-based sex estimates. J Forensic Sci 67: 68-79.

Hastie T, Tibshirani R, Friedman J. 2009. The elements of statistical learning: Data mining, inference and prediction. 2nd ed. New York: Springer-Verlag.

Hauser G, de Stefano GF. 1989. Epigenetic variants of the human skull. Schweizerbart, Stuttgart.

Hefner JT. 2003. Assessing nonmetric cranial traits currently used in forensic determination of ancestry [PhD dissertation]. University of Florida.

Hefner JT, Ousley SD, Warren M. 2004. An historical perspective on nonmetric skeletal variation: Hooton and the Harvard list. In: Proceedings of the 56th Annual Meeting of the American Academy of Forensic Sciences. p. 16-21.

Hefner JT. 2009. Cranial Nonmetric Variation and Estimating Ancestry. J Forensic Sci 54:985-995.

Hefner JT, Ousley SD, Dirkmaat DC. 2012. Morphoscopic traits and the assessment of ancestry. In: Dirkmaat DC (ed). A companion to Forensic Anthropology. Blackwell Publishing LTD. p.287-310.

Hefner JT, Ousley SD. 2014. Statistical classification methods for estimating ancestry using morphoscopic traits. J Forensic Sci 59:883-890.

Hefner JT, Spradley MK, Anderson B. 2014. Ancestry assessment using random forest modeling. J Forensic Sci 59:583-589.

Hefner JT, Pilloud MA, Black CJ, Anderson BE. 2015. Morphoscopic trait expression in "Hispanic" populations. J Forensic Sci 60: 1135-1139.

Hefner JT. 2018. The macromorphoscopic databank. Am J Phys Anthropol 166: 994-1004.

Hefner JT, Plemons A, Kamnikar KR, Ousley SD, Linde K. (n.d.). MMS v1.61: User Manual v. 1.0. Retrieved from http://macromorphoscopic.com/mms-software/

Hefner, J. (2020). MaMD Analytical. Retrieved from https://github.com/rer145/mamd-analytical

Huang Z, Chen H, Hsu CJ, Chen WH, Wu S. 2004. Credit rating analysis with support vector machines and neural networks: A market comparative study. Decis Support Syst 37: 543-558.

Hufnagl HBD. 2015. Age estimation with decision trees: Testing the relevance of 94 aging indicators on the William M. Bass donated collection [PhD dissertation]. University of Tennessee.

İşcan MY, Steyn M. 1999. Craniometric determination of population affinity in South Africans. Int J Legal Med 112:91-97.

James G, Witten D, Hastie T, Tibshirani R. 2013. An Introduction to Statistical Learning. Springer, New York.

Jantz RL, Ousley SD. 2005. FORDISC 3.0: Personal Computer Forensic Discriminant Functions. Knoxville: University of Tennessee.

Jantz RL, Ousley SD. n.d. FORDISC 3.0: Personal Computer Forensic Discriminant Functions: Help File (version 1.53) Knoxville: University of Tennessee.

Kamnikar KR, Plemons AM, Hefner JT. 2018. Intraobserver error in macromorphoscopic trait data. J Forensic Sci 63:361-370.

Kimmie-Dhansay F, Pontes CC, Chikte U, Erasmus RT, Kengne AP, Matsha TE. 2021. Tooth loss in relation to serum cotinine levels - A cross-sectional study from the Belville South area in South Africa. S Afr Dent J 76: 207-215.

Klales AR, Ousley SD, Vollner JM. 2012. A revised method of sexing the human innominate using Phenice's nonmetric traits and statistical methods. Am J Phys Anthropol 149: 104-114.

Klales AR, Kenyhercz MW. 2015. Morphological assessment of ancestry using cranial macromorphoscopics. J Forensic Sci 60:13-20.

Klales AR. 2020. MorphoPASSE: Morphological pelvis and skull sex estimation program. In: Sex estimation of the human skeleton. p. 271-278. Academic Press.

Klales AR. 2021. Current State of Sex Estimation in Forensic Anthropology. J Forensic Anthropol 4: 118.

Klales AR, Garvin H, Gocha TP, Lesciotto KM, Walls M. 2020. Examining the Reliability of Morphological Traits for Sex Estimation: Implications for the Walker (2008) and Klales et al. (2012) Methods. J Forensic Anthropol 3: 139-150.

Klepinger L. 2006. Fundamentals of forensic anthropology. John Wiley and Sons, Inc. p. 64-76.

Krogman WM. 1962. The human skeleton in forensic medicine. Charles C. Thomas: Springfield, IL.

Krüger GC. 2015. Comparison of sexually dimorphic patterns in the postcrania of South Africans and North Americans [Masters dissertation]. University of Pretoria.

Krüger GC. L'Abbé EN, Stull KE, Kenyhercz MW. 2015. Sexual dimorphism in cranial morphology among modern South Africans. Int J Legal Med 129:869-875.

Krüger GC, L'Abbé EN, Stull KE. 2017. Sex estimation from the long bones of modern South Africans. Int J Legal Med 131:275-285.

Krüger GC, Liebenberg L, Myburgh J, Meyer A, Oettlé AC, Botha D, Brits DM, Kenyhercz MW, Stull KE, Sutherland C, L'Abbé EN. 2018. Forensic Anthropology and the Biological Profile in South Africa. In: Latham K, Bartelink E, Finnegan M (eds). New Perspectives in Forensic Human Skeletal Identification. Elsevier Academic Press. p. 313-321.

Kuzminsky SC, Snyder TJ, Tung TA. 2020. The limited efficacy of 3D models for teaching students sex estimations based on cranial traits: A case for investment in osteology teaching labs. Int J Osteoarchaeol 30: 275-280.

L'Abbé EN, Loots M, Meiring JH. 2005. The Pretoria Bone Collection: A modern South African skeletal sample. Homo 56:197-205.

L'Abbé EN, van Rooyen C, Nawrocki SP, Becker PJ. 2011. An evaluation of non-metric cranial traits used to estimate ancestry in a South African sample. Forensic Sci Int 209:195-e1.

L'Abbé EN, Kenyhercz MW, Stull KE, Keough N, Nawrocki S. 2013. Application of Fordisc 3.0 to explore differences among crania of North American and South African blacks and whites. J Forensic Sci 6:1579-1583.

Landis JR, Koch GG. 1977. The measurement of observer agreement for categorical data. Biometrics: 159-174.

Langley NR, Jantz LM, McNulty S, Maijanen H, Ousley SD, Jantz RL. 2018. Error quantification of osteometric data in forensic anthropology. Forensic Sci Int 287: 183-189.

Lee SW. 2022. Methods for testing statistical differences between groups in medical research: statistical standard and guideline of Life Cycle Committee. Life Cycle 2: e1.

Lewis CJ, Garvin HM. 2016. Reliability of the Walker cranial nonmetric method and implications for sex estimation. J Forensic Sci 61: 743-751.

Liaw A, Wiener M. 2002. randomForest: Classification and Regression by randomForest. Retrieved from: https://CRAN.R-project.org/package=randomForest

Liebenberg L, L'Abbé EN, Stull KE. 2015a. Population differences in the postcrania of modern South Africans and the implications for ancestry estimation. Forensic Sci Int 257:522-529.

Liebenberg L, Stull KE, L'Abbé EN, Botha D. 2015b. Evaluating the accuracy of cranial indices in ancestry estimation among South African groups. J Forensic Sci 60: 1277-1282.

Liebenberg L, Krüger GC, L'Abbé EN, Stull KE. 2019. Postcraniometric sex and ancestry estimation in South Africa: A validation study. Int J Legal Med: 1-8.

Liebenberg L, Krüger GC. 2020. Standardization and quality assurance in skeletal landmark placement and osteometry. Forensic Sci Int 308: 110168.

Lieberman L, Kirk RC, Corcoran M. 2003. The decline of race in American physical anthropology. Przegl Anthropol 66: 3-21.

Littlefield A, Lieberman L, Reynolds LT. 1982. Redefining race: The potential demise of a concept in physical anthropology. Curr Anthropol 23: 641-655.

Maass P, Friedling LJ. 2019. Morphometric analysis of the neurocranium in an adult South African cadaveric sample. J Forensic Sci 64: 367-374.

Maier CA. 2017. The Combination of Cranial Morphoscopic and Dental Morphological Methods to Improve the Forensic Estimation of Ancestry [PhD dissertation]. University of Nevada, Reno.

Maier CA. 2019. Evaluating Mixed-Methods Models for the Estimation of Ancestry from Skeletal Remains. J Forensic Anthropol 2: 45-56.

Maier CA, George RL. 2020. Examining differences in presumed migrants from Texas and Arizona using cranial and dental data. J Forensic Anthropol 3: 17-28.

Martínez-Abadías N, Esparza M, Sjøvold T, González-José R, Santos M, Hernández M. 2009. Heritability of human cranial dimensions: comparing the evolvability of different cranial regions. J Anat 214: 19-35.

Martínez-Abadías N, Mitteroecker P, Parsons TE, Esparza M, Sjøvold T, Rolian C, Richtsmeier JT, Hallgrímsson B. 2012. The developmental basis of quantitative craniofacial variation in humans and mice. Evol Biol 39: 554-567.

McDowell JL, L'Abbé EN, Kenyhercz MW. 2012. Nasal aperture shape evaluation between black and white South Africans. Forensic Sci Int 222: 397.e1-397.e6.

McDowell JL, Kenyhercz MW, L'Abbé EN. 2015. An evaluation of nasal bone and aperture shape among three South African populations. Forensic Sci Int 252: 189-e1.

McHugh ML. 2012. Interrater reliability: the kappa statistic. Biochem Med 22: 276-282.

Meintjes-Van der Walt L. 2003. The proof of the pudding: The presentation and proof of expert evidence in South Africa. J Afr Law 47:88-106.

Merchant C. 2023. Ancestry estimates: Evaluating the reliability of Hefner's cranial morphoscopic method. [MA dissertation]. University of Manitoba.

Michael A, Bengtson J, Blatt S. 2021. Genes, race, ancestry, and identity in forensic anthropology: Historical perspectives and contemporary concerns. Forensic Genomics 1: 41-46.

Mitteroecker P, Gunz P, Neubauer S, Müller G. 2012. How to explore morphological integration in human evolution and development? Evol Biol 39: 536-553.

Moore-Jansen PM, Ousley SD, Jantz RL. 1994. Data collection procedures for forensic skeleton material. The University of Tennessee, Knoxville: Department of Anthropology. p. 70-82.

Navega DL, Coelho C, Vicente R, Ferreira MT, Wasterlain S, Cunha E. 2015. AncesTrees: Ancestry estimation with randomized decision trees. Int J Legal Med 129:1145-1153.

Nikita E. 2014. Age-associated variation and sexual dimorphism in adult cranial morphology: Implications in anthropological studies. Int J Osteoarchaeol 24: 557-569.

Oettlé AC, Steyn M. 2000. Age estimation from sternal ends of ribs by phase analysis in South African blacks. J Forensic Sci 45: 1071-1079.

Ossenberg NS. 1976. Within and between race distances in population studies based on discrete traits of the human skull. Am J Phys Anthropol 45: 701-715.

Ousley SD, Jantz RL, Freid D. 2009. Understanding Race and Human Variation: Why Forensic Anthropologists are good at Identifying Race. Am J Phys Anthropol 139:68-75.

Ousley SD. 2016. Forensic classification and biodistance in the 21st century: The rise of learning machines. In: Pilloud MA, Hefner JT (eds). Biological Distance Analysis. p. 197-212. Academic Press.

Parr NM. 2005. Determination of ancestry from discrete traits of the mandible [PhD dissertation]. University of Indianapolis.

Patterson E, Sethuram A, Albert AM, Ricanek Jr. K, King M. 2007. Aspects of age variation in facial morphology affecting biometrics. In: 2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems. pp. 1-6.

Perini TA, de Oliveira GL, Ornellas JDS, de Oliveira FP. 2005. Technical error of measurement in anthropometry. Rev Bras Med Esporte 11:81-85.

Pink CM, Maier C, Pilloud MA, Hefner JT. 2016. Cranial nonmetric and morphoscopic data sets. In: Pilloud MA, Hefner JT (eds). Biological Distance Analysis: Forensic and Bioarchaeological perspectives. Elsevier Inc. p. 91-107.

Plemons AM, Hefner JT. 2016. Ancestry estimation using macromorphoscopic traits. Acad Forensic Pathol 6:400-412.

Plemons AM, Hefner JT, Kamnikar KR. 2018. Refining Asian ancestry classifications via cranial macromorphoscopic traits. Am J Phys Anthropol 165: 210.

President's Council of Advisors on Science and Technology (PCAST). 2016. Report to the president Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Washington DC.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Relethford JH. 1994. Craniometric variation among modern human populations. Am J Phys Anthropol 95: 53-62.

Relethford JH. 2002. Apportionment of global human genetic diversity based on craniometrics and skin color. Am J Phys Anthropol 118: 393-398.

Rhine S. 1990. Non-metric skull racing. In: Gill GW and Rhine S (Eds). Skeletal attribution of race: Methods for forensic anthropology. University of New Mexico, Albuquerque. p. 9-20.

Ross AH, Williams SE. 2010. Craniofacial growth, maturation, and change: teens to midadulthood. J Craniofac Surg 21: 458-461.

Ross AH, Pilloud M. 2021. The need to incorporate human variation and evolutionary theory in forensic anthropology: A call for reform. Am J Phys Anthropol 176: 672-683.

Sauer N. 1992. Forensic anthropology and the concept of race: If races don't exist, why are forensic anthropologists so good at identifying them? Soc Sci Med 34:107-111.

Scariano SM, Davenport JM. 1987. The effects of violations of independence assumptions in the one-way ANOVA. Am Stat 41: 123-129.

Shirley NR, Ramirez Montes PA. 2015. Age estimation in forensic anthropology: quantification of observer error in phase versus component-based methods. J Forensic Sci 60: 107-111.

Sim J, Wright CC. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 85: 257-268.

Slice DE. 2007. Geometric morphometrics. Annu Rev Anthropol 36: 261-281.

Small C, Brits D, Hemingway J. 2016. Assessing the effects of tooth loss in adult crania using geometric morphometrics. Int J Legal Med 130: 233-243.

Small C, Schepartz L, Hemingway J, Brits D. 2018. Three-dimensionally derived interlandmark distances for sex estimation in intact and fragmentary crania. Forensic Sci Int 287: 127-135.

Smith AC, Boaks A. 2017. Consistency of selected craniometric landmark locations and the resulting variation in measurements. Forensic Sci Int 280: 156-163.

Smithsonian Institution. 2012. Osteoware: Standardized Skeletal Documentation Software.

Spradley M K, Jantz RL. 2011. Sex Estimation in Forensic Anthropology: Skull Versus Postcranial Elements. J Forensic Sci 56: 289-296.

Spradley MK, Jantz RL. 2016. Ancestry estimation in forensic anthropology: Geometric morphometric versus standard and non-standard interlandmark distances. J Forensic Sci 61: 892-897.

Spradley MK, Jantz RL. 2021. What are we really estimating in forensic anthropological practice, population affinity or ancestry? J Forensic Anthropol 4: 309-318.

Spradley MK, Stull KE. 2018. Advancements in sex and ancestry estimation. In: Latham K, Bartelink E, Finnegan M (eds). New Perspectives in Forensic Human Skeletal Identification. Elsevier Academic Press. p. 13-21.

Statistics South Africa. 2022. Mid-year population estimates: Statistical Release.

Steyn M, İşcan MY. 1998. Sexual dimorphism in the crania and mandibles of South African whites. Forensic Sci Int 98: 9-16.

Stomfai S, Ahrens W, Bammann K, Kovács É, Mårild S, Michels N, Moreno LA, Pohlabeln H, Siani A, Tornaritis M, Veidebaum T, Molnár D. 2011. Intra- and inter-observer reliability in anthropometric measurements in children. Int J Obesity 35:45-51.

Strobl C, Malley J, Tutz G. 2009. An Introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. Psychol Methods 14:323.

Stull KE, Kenyhercz MW, L'Abbé EN. 2014a. Ancestry estimation in South Africa using craniometrics and geometric morphometrics. Forensic Sci Int 245:206.

Stull KE, Tise ML, Ali Z, Fowler DR. 2014b. Accuracy and reliability of measurements obtained from computed tomography 3D volume rendered images. Forensic Sci Int 238: 133-140.

Stull KE, Kenyhercz MW, Tise ML, L'Abbé EN, Tuamsuk P. 2016. The craniometric implications of a complex population history in South Africa. In: Pilloud MA, Hefner JT (eds). Biological Distance Analysis: Forensic and Bioarchaeological perspectives. Elsevier Inc. p. 245-263.

Stull KE, Bartelink EJ, Klales AR, Berg GE, Kenyhercz MW, L'Abbé EN, Go MC, McCormick K, Mariscal C. 2021. Commentary on: Bethard JD, DiGangi EA. Letter to the Editor-Moving beyond a lost cause: Forensic anthropology and ancestry estimates in the United States. J Forensic Sci. 2020; 65 (5): 1791-2. J Forensic Sci 66: 417-420.

Tabachnick BG, Fidell LS. 2007. Using multivariate statistics. 5th ed. Boston: Pearson Education.

Tallman SD, Parr NM, Winburn AP. 2021. Assumed Differences; Unquestioned Typologies: The Oversimplification of Race and Ancestry in Forensic Anthropology. J Forensic Anthropol 4: 73 – 96.

Taylor R. 1990. Interpretation of the correlation coefficient: A basic review. J Diagn Med Sonogr 6:35-39.

Tishkoff SA, Williams SM. 2002. Genetic analysis of African populations: Human Evolution and complex disease. Nat Rev Genet 3:611-621.

Tobias PV. 1985. History of physical anthropology in southern Africa. Am J Phys Anthropol 28: 1-52.

Tran D, Dolgun A, Demirhan H. 2020. Weighted inter-rater agreement measures for ordinal outcomes. Commun Stat Simul 49: 989-1003.

Viðarsdóttir US, O'Higgins P, Stringer C. 2002. A geometric morphometric study of regional differences in the ontogeny of the modern human facial skeleton. J Anat 201: 211-229.

von Cramon-Taubadel N, Frazier BC, Lahr MM. 2007. The problem of assessing landmark error in geometric morphometrics: Theory, methods and modifications. Am J Phys Anthropol 134:24-35.

von Cramon-Taubadel N. 2014. Evolutionary insights into global patterns of human cranial diversity: Population history, climatic and dietary effects. J Anthropol Sci 91:1-36.

Walker PL. 2008. Sexing skulls using discriminant function analysis of visually assessed traits. Am J Phys Anthropol: 36: 39-50.

Walrath DE, Turner P, Bruzek J. 2004. Reliability test of the visual assessment of cranial traits for sex determination. Am J Phys Anthropol 125:132-137.

Weinberg SM, Putz DA, Mooney MP, Siegel MI. 2005. Evaluation of non-metric variation in the crania of black and white perinates. Forensic Sci Int 151: 177-185.

Wilczak CA, Mariotti V, Pany-Kucera D, Villotte S, Henderson CY. 2017. Training and interobserver reliability in qualitative scoring of skeletal samples. J Archaeol Sci: Reports 11: 69-79.

Williams SE, Slice DE. 2010. Regional shape change in adult facial bone curvature. Am J Phys Anthropol 143: 437 – 447.

Wood C. 2015. The age-related emergence of cranial morphological variation. Forensic Sci Int 251: 220-e1.

APPENDIX

Appendix I – Macromorphoscopic trait definitions and states

Trait name	State	Definition
Anterior nasal spine **(ANS)**	**Taken in the area of the nasal spine, best viewed from anterior aspect.**	
	1	Minimal to no projection of spine beyond inferior nasal aperture.
	2	Moderate projection of spine beyond inferior nasal aperture.
	3	Pronounced projection of spine beyond inferior nasal aperture.
Inferior nasal aperture **(INA)**	**Refers to shape of border which defines transition from nasal floor to vertical potion of the maxilla.**	
	1	Inferior sloping begins within nasal cavity and terminates on vertical surface of maxilla; produces smooth transition.
	2	Sloping begins more anteriorly, with more angulation at the exit of the nasal opening.
	3	Transition from nasal floor to vertical maxilla is not sloping, nor is there an intervening sill (i.e. forms a right angle).
	4	Superior incline of the anterior nasal floor creates a weak (but present) vertical ridge traversing the border (partial sill).
	5	Pronounced ridge obstructing the nasal floor-to-maxilla transition (sill).
Inter-orbital breadth **(IOB)**	**Assessment made relative to facial skeleton.**	
	1	Narrow.
	2	Medium.
	3	Broad.
Malar tubercle **(MT)**	**Caudally protruding tubercle on inferior margin of maxilla/zygoma in the region of the zygomaticomaxillary suture. Score with a transparent ruler.**	
	0	No projection.
	1	Trace tubercle about 2mm or less below ruler's edge.
	2	Medium protrusion about 2mm - 4mm below ruler's edge.
	3	Pronounced tubercle about 4mm or more below ruler's edge.
Nasal aperture shape **(NAS)**	**Assessed by observing lateral contours of nasal aperture and the position of the greatest lateral projection of the margin.**	
	1	Teardrop; greatest lateral projection intermediate to 2 and 3.
	2	Bell-shaped; greatest projection at inferior margin.
	3	Bowed; greatest projection at midline.

Nasal aperture width **(NAW)**	**Assessment made relative to facial skeleton.**	
	1	Narrow.
	2	Medium.
	3	Broad.
Nasal bone contour **(NBC)**	**Contour of nasal bones and frontal process of maxilla taken 1cm below nasion. Scored with a contour gauge.**	
	0	Low and rounded contour (circular shape, lacks steep walls).
	1	Oval contour (elongated with high and rounded lateral walls).
	2	Broad plateau (steep walls with broad flat superior surface).
	3	Narrow plateau (steep walls with narrow flat superior surface).
	4	Triangular (steep walls lacking a flat superior surface).
Nasal bone shape **(NBS)**	**Assess lateral contours of nasal bones rather than the frontonasal suture, nasal suture or symmetry of the nasal bones.**	
	1	No nasal pinch (bones may be wide or narrow).
	2	Superior pinch with minimal lateral bulging.
	3	Superior pinch with pronounced lateral bulging inferiorly.
	4	Triangular.
Nasal overgrowth **(NO)**	**Inferior projection of lateral borders of the nasal bones beyond the maxillae. Does not include anterior bulging of nasal bones.**	
	0	No overgrowth.
	1	Any projection at nasale inferius present.
Nasofrontal suture **(NFS)**	**Suture separating nasal bones from the frontal bone. Ignore symmetry of the nasal bones.**	
	1	Round and lacks angles.
	2	Square (approximate right angles at nasale superius).
	3	Triangular.
	4	Irregular (lacks any definite shape).
Orbital shape **(OS)**	**Defined by the shape of the orbital borders, best scored from anterior view.**	
	1	Rectangular (horizontal margins longer than vertical margins).
	2	Circular (margins equidistant from centre on all sides).
	3	Rhombic (medial border shorter than the lateral border; “aviator sunglasses”).
	Found along sagittal suture, observed in lateral profile.	
	0	No depression.

Post-bregmatic depression **(PBD)**	**1**	Marked depression.
Posterior zygomatic tubercle **(PZT)**	**Marginal process; posterior projection of the zygoma at midorbit. Score with a transparent ruler.**	
	0	No projection.
	1	Weak projection (less than 4mm).
	2	Moderate projection (4 – 6 mm).
	3	Marked projection (more than 6mm).
Supranasal suture **(SPS)**	**Persistent, complex suture represents fusion of the nasal portion of the frontal suture; it is not a persistent metopic suture.**	
	0	Completely obliterated.
	1	Open (unfused).
	2	Closed, but visible.
Transverse palatine suture **(TPS)**	**View entire suture (i.e. not unilateral), concentrating on the medial section.**	
	1	Courses perpendicular to median palatine suture with no anterior/posterior deviations.
	2	Courses perpendicular to median palatine suture with an anterior bulge/deviation.
	3	Deviates anteriorly or posteriorly in the region of the median palatine suture (similar to EKG reading).
	4	Courses perpendicular to median palatine suture with a posterior bulge/deviation.
Palate shape **(PS)**	**Contour of the dental arcade, viewed from inferior surface.**	
	1	Elliptical (smooth round curvature with a constriction at the area of the third molars)
	2	Parabolic A (smooth round curvature with flaring at the area of the third molars)
	3	Parabolic B (similar to 2, but palate is longer than it is wide)
	4	Hyperbolic (slightly flattened, parallel configuration; square)
Zygomatico-maxillary suture **(ZS)**	**Based primarily on location of greatest lateral projection and the number of angles present.**	
	0	No angles, greatest projection inferiorly at the malar tubercle.
	1	One angle, greatest projection near midline.
	2	Two or more angles (s-shaped), greatest projection is variable.

Appendix II – Measurement definitions

Maximum cranial length (GOL) – The straight, maximum distance between glabella and opisthocranion (spreading caliper).

Maximum cranial breadth (XCB) – The maximum width of the skull from euryon to euryon, taken perpendicular to the midsagittal plane (spreading caliper).

Bizygomatic breadth (ZYB) – The maximum direct distance between the most lateral points on the zygomatic arches, from zygion to zygion (spreading caliper).

Basion-bregma height (BBH) – The direct distance in the midplane between basion and bregma. Mark basion with a pencil for greater repeatability (spreading caliper).

Basion-nasion length (BNL) – The direct distance in the midplane between basion and nasion. Mark basion with a pencil for greater repeatability (spreading caliper).

Basion-prosthion length (BPL) – The direct distance in the midplane between basion and prosthion. Mark basion with a pencil for greater repeatability. Do not take if specimen is edentulous (spreading caliper).

Maximum alveolar length (MAL) – The direct distance in the midplane between prosthion and alveolon. Do not take if specimen is edentulous (sliding caliper).

Maximum alveolar breadth (MAB) – The maximum breadth on the lateral border of the maxilla at the location of the second molars. Do not take if specimen is edentulous (sliding caliper).

Biasterionic breadth (ASB) – The direct distance between the left and right asterion (sliding caliper).

Upper facial height (NPH) – The direct distance in the midplane between nasion and prosthion. Do not take if specimen is edentulous (sliding caliper).

Minimum frontal breadth (WFB) – The minimum distance between the left and right frontotemporale (sliding caliper).

Upper facial breadth (UFBR) – The direct distance between the most lateral points on the frontomalar suture (sliding caliper).

Nasal height (NLH) – The average height from nasion to the lowest point on the border of the nasal aperture on either side (sliding caliper).

Nasal breadth (NLB) – The maximum breadth of the nasal aperture taken at the most lateral points on the margin (sliding caliper).

Orbital breadth (OBB) – The laterally sloping distance from dacryon to ectoconchion. Mark ectoconchion with a pencil for greater repeatability (sliding caliper).

Orbital height (OBH) – The distance between the superior and inferior margins of the orbit perpendicular to the orbital breadth and bisecting the orbit into equal halves (sliding caliper).

Bi-orbital breadth (EKB) – The direct distance between the left and right ectoconchion. Mark ectoconchion with a pencil for greater repeatability (sliding caliper).

Inter-orbital breadth (DKB) – The direct distance between the left and right dacryon (sliding caliper).

Frontal chord (FRC) – The direct distance between nasion and bregma (sliding caliper).

Parietal chord (PAC) – The direct distance between bregma and lambda (sliding caliper).

Occipital chord (OCC) – The direct distance between lambda and opisthion (sliding caliper).

Foramen magnum length (FOL) – The direct distance between basion and opisthion. Mark basion with a pencil for greater repeatability (sliding caliper).

Foramen magnum breadth (FOB) – The distance between the lateral margins of the foramen magnum at the point of greatest curvature (sliding caliper).

Mastoid height (MDH) – The direct distance between porion and mastoidale (sliding caliper).

Bi-auricular breadth (AUB) – The minimum distance across the roots of the zygomatic processes, from auriculare to auriculare (spreading caliper).

Appendix III – Shapiro Wilk test for normality of metric data

Variable	p-value
GOL	0.48
XCB	0.05
ZYB	**<0.01**
BBH	**0.04**
BNL	**0.01**
BPL	0.11
MAL	0.15
MAB	0.05
ASB	**0.01**
NPH	0.05
WFB	**<0.01**
UFBR	**0.01**
NLH	**<0.01**
NLB	**<0.01**
OBB	**<0.01**
OBH	**<0.01**
EKB	**<0.01**
DKB	**<0.01**
FRC	**0.01**
PAC	0.05
OCC	**0.01**
FOL	**<0.01**
FOB	**<0.01**
MDH	**<0.01**
AUB	0.15

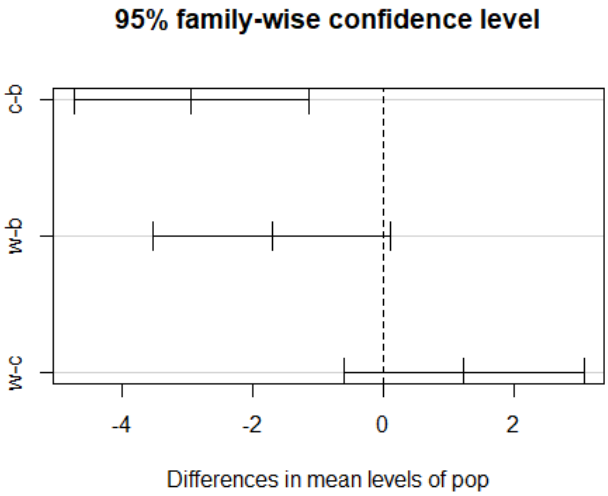
*** Bold indicates deviation from normal distribution (p<0.05)**

Appendix IV – Levene's test for homoscedasticity of variance for metric data

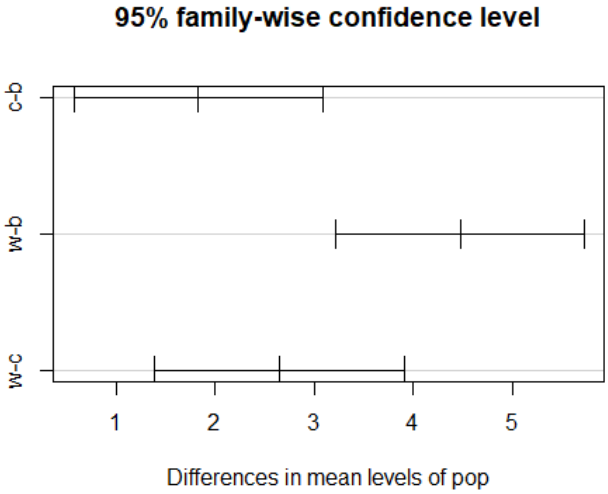
Variable	p-value
GOL	0.16
XCB	0.56
ZYB	0.40
BBH	0.43
BNL	0.91
BPL	0.21
MAL	0.35
MAB	0.68
ASB	0.82
NPH	**0.04**
WFB	**0.04**
UFBR	0.55
NLH	0.48
NLB	0.39
OBB	0.95
OBH	**<0.01**
EKB	0.43
DKB	0.84
FRC	0.76
PAC	0.86
OCC	0.14
FOL	0.69
FOB	0.51
MDH	0.21
AUB	**0.04**

*** Bold indicates unequal variance-covariance matrices ($p<0.05$)**

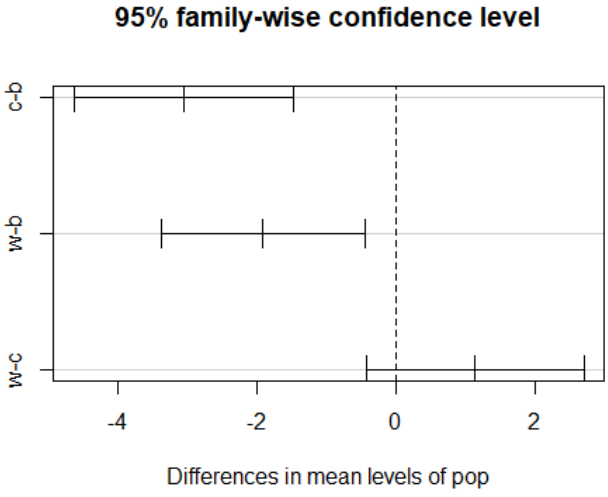
Appendix V – Tukey's HSD for metric variables



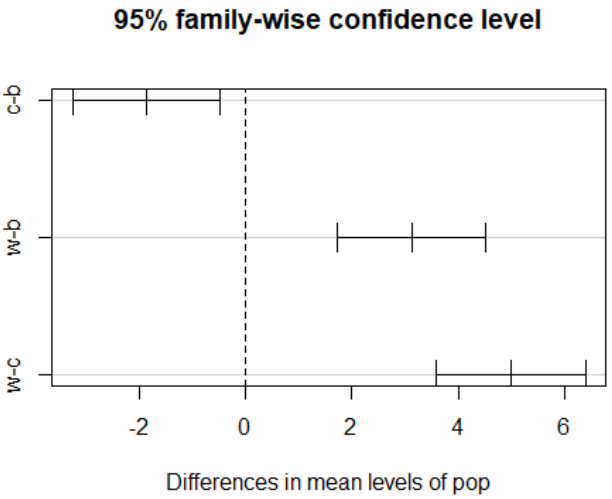
GOL



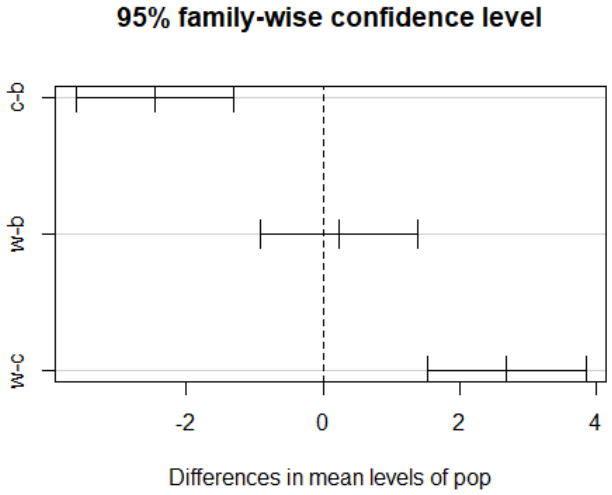
XCB



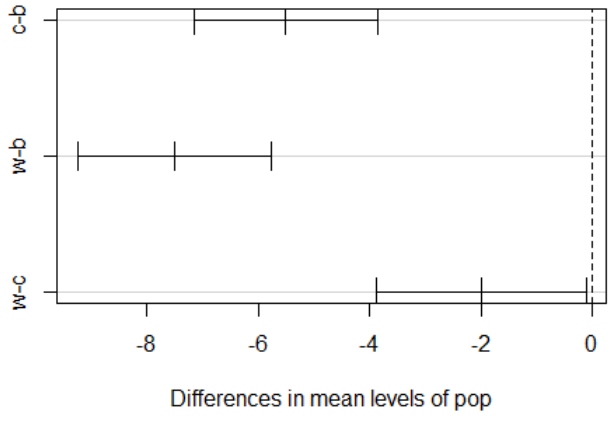
ZYB



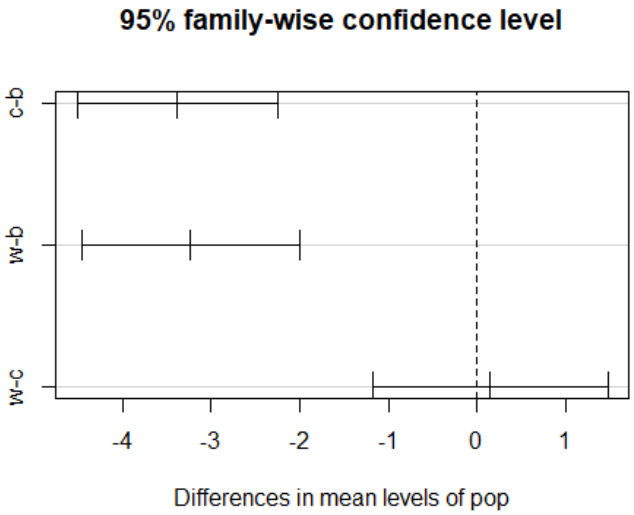
BBH



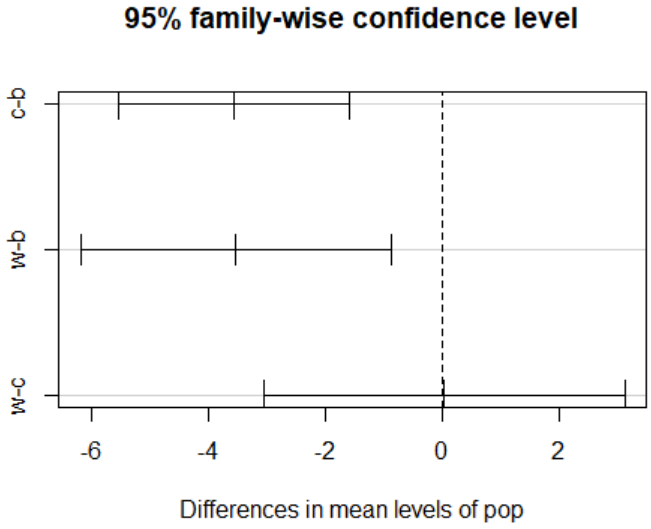
BNL



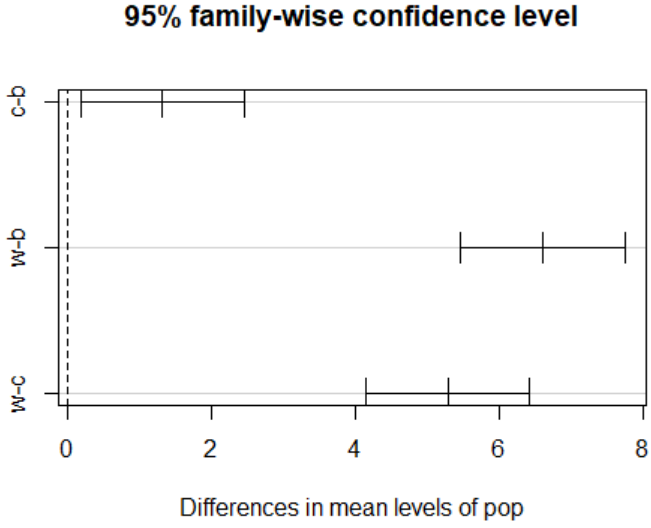
BPL



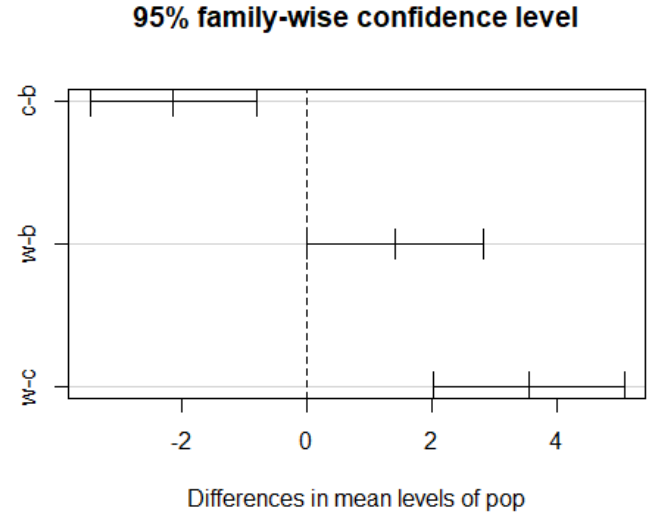
MAL



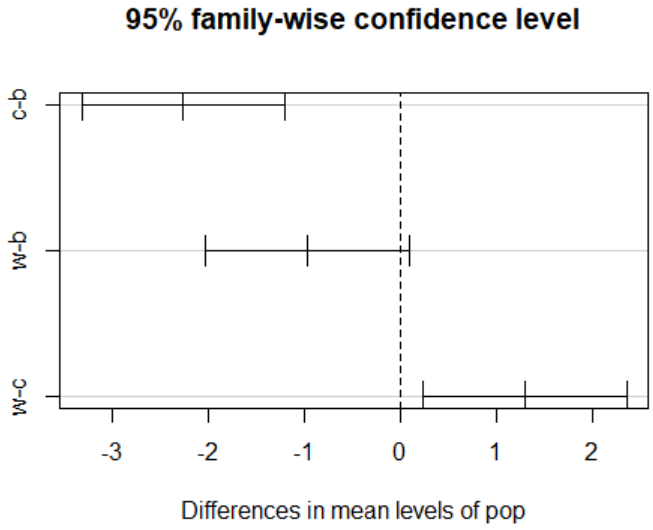
MAB



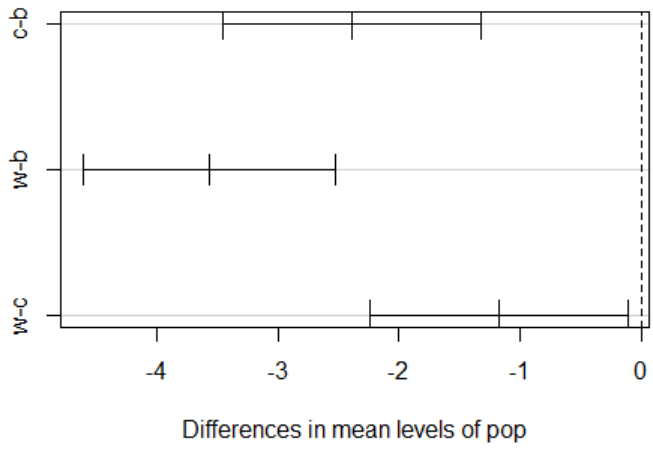
ASB



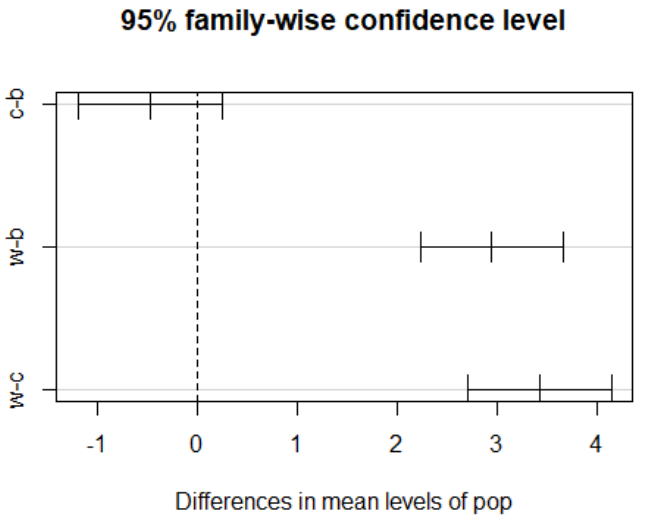
NPH



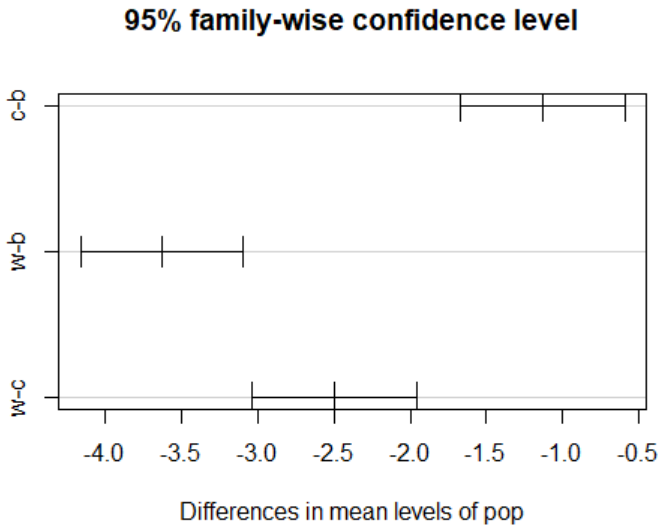
WFB



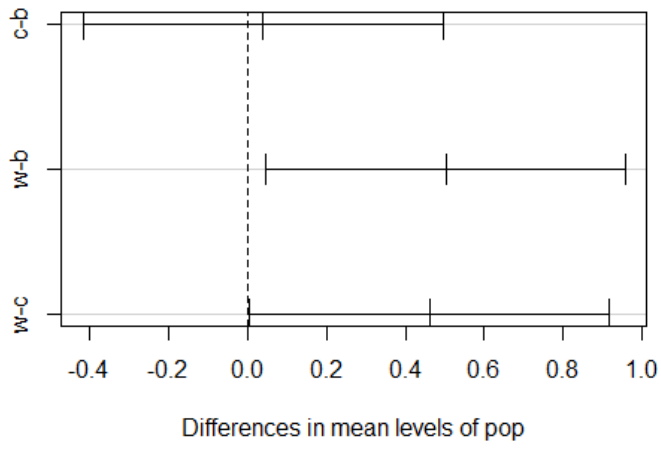
UFBR



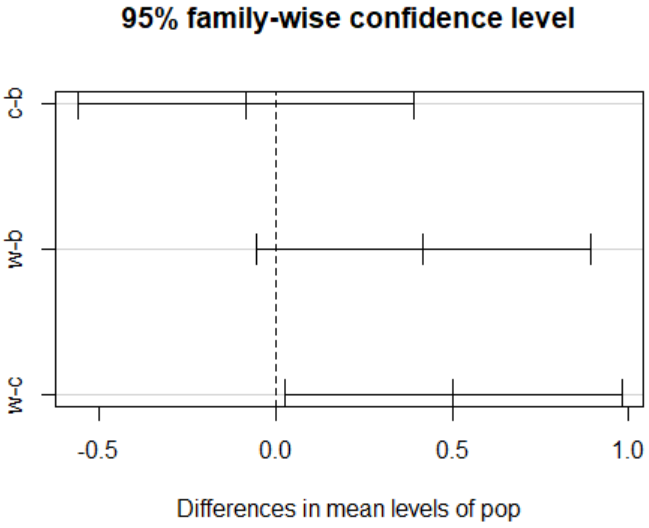
NLH



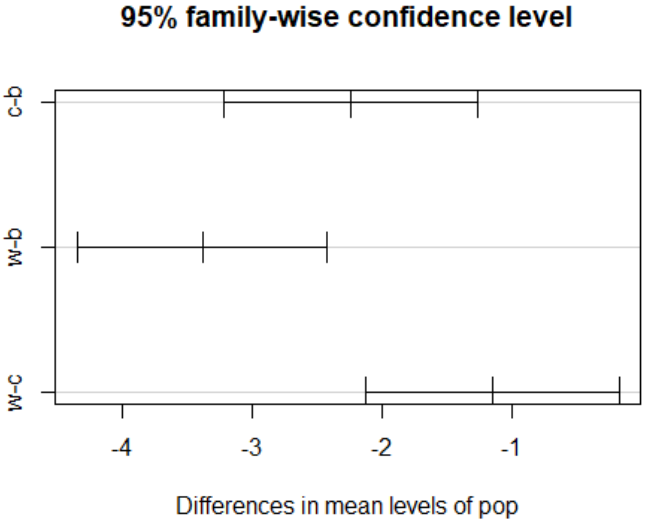
NLB



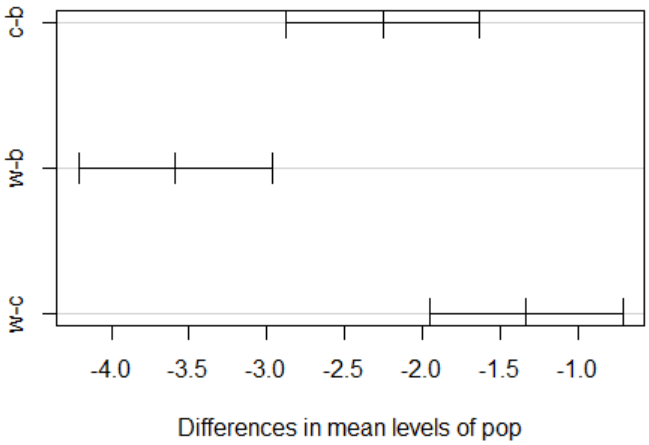
OBB



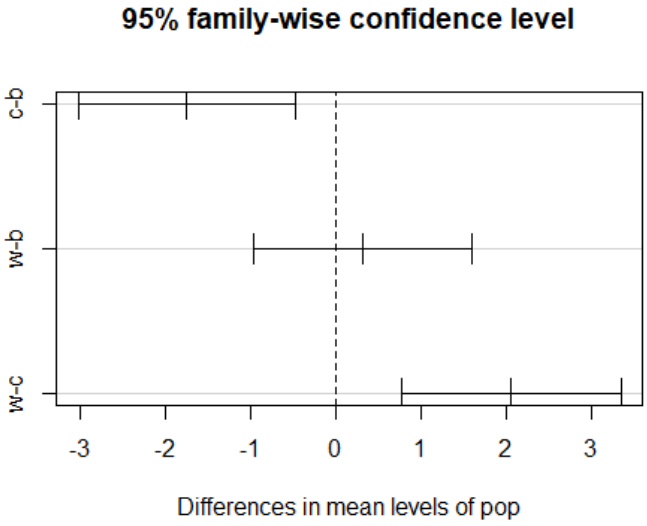
OBH



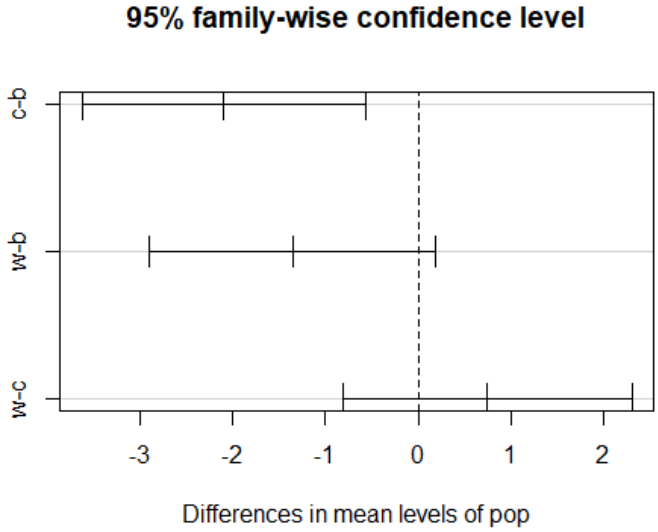
EKB



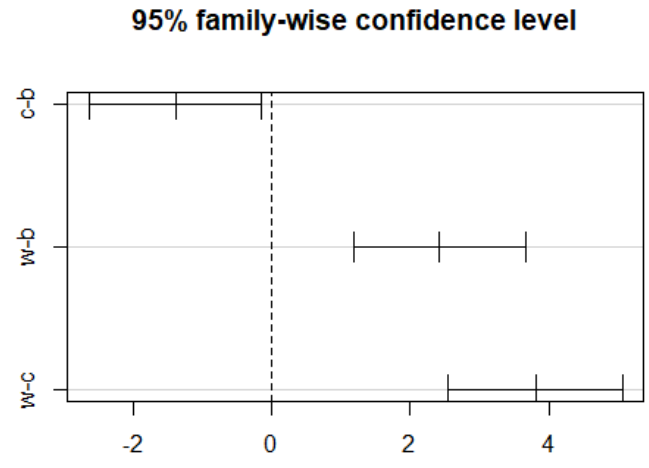
DKB



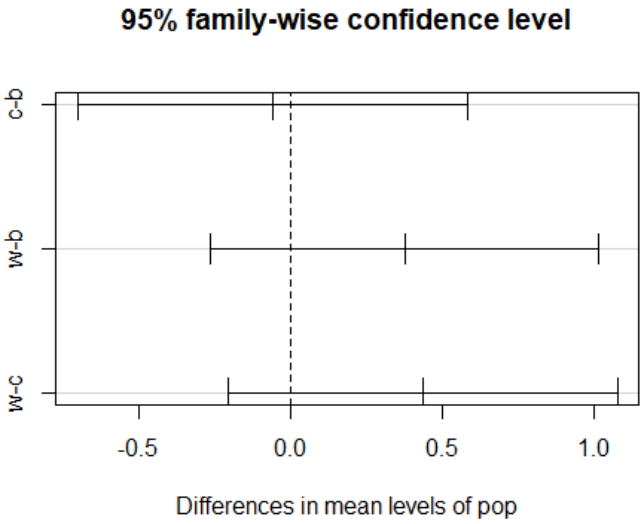
FRC



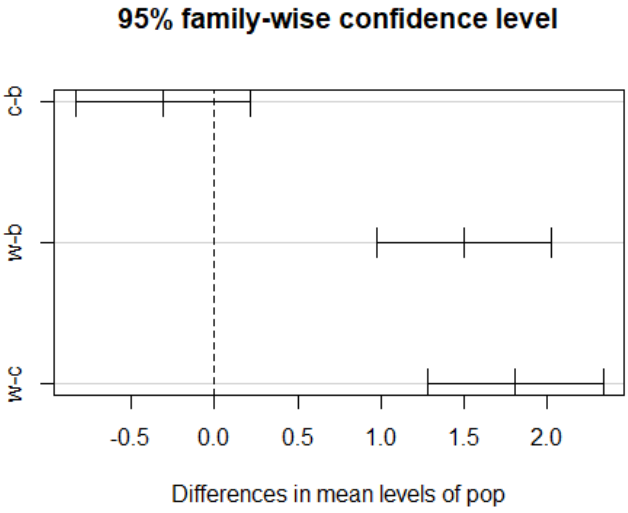
PAC



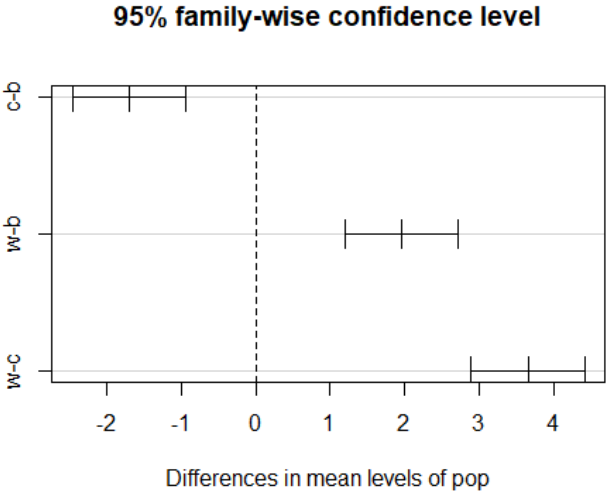
OCC



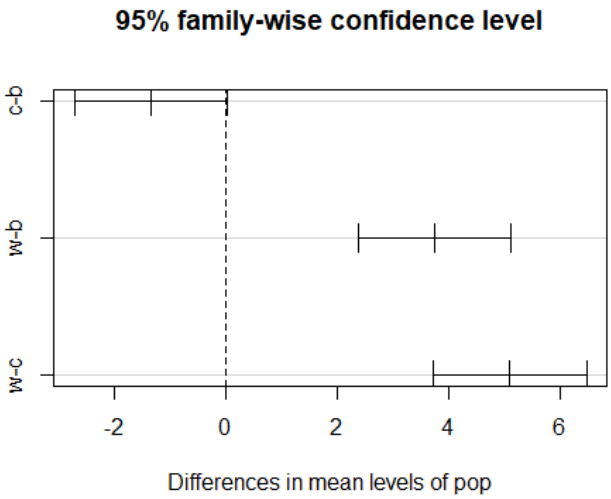
FOL



FOB



MDH



AUB

Appendix VI - Comparison of published observer error rates for MMS traits

	Intra-observer					Inter-observer				
	Current study	**Hefner (2009)**	**L'Abbé et al. (2011)**	**Maier (2017)**	**Kamnikar et al. (2018)**	**Current study**	**Hefner (2009)**	**L'Abbé et al. (2011)**	**Klales and Kenyhercz (2014) ***	**Klales and Kenyhercz (2014) ****
ANS	0.821	0.422	0.81	0.759	0.49	0.685	0.506	0.55	0.165	-0.250
INA	0.468	0.964	0.58	-	0.63	0.357	0.376	0.65	0.284	-0.522
IOB	0.833	0.857	0.53	0.666	0.64	0.689	0.325	0.44	0.412	0.242
MT	0.717	0.929	0.53	0.658	0.10	0.551	0.470	0.44	0.382	-0.538
NAS	0.615	-	-	-	-	0.356	-	-	0.324	0.412
NAW	0.906	0.929	0.68	0.702	0.64	0.752	0.732	0.56	0.167	0.167
NBC	0.643	0.810	0.74	0.624	0.69	0.159	0.231	0.54	0.141	0.032
NBS	0.429	-	-	-	-	0.297	-	-	0.198	0.155
NO	0.412	1.00	0.64	0.922	0.58	0.695	1.00	0.73	0.374	0.500
NFS	0.833	-	-	-	-	0.506	-	-	0.281	0.032
OS	0.804	-	-	-	-	0.464	-	-	0.453	0.375
PBD	0.737	0.820	-	0.768	0.62	0.255	0.232	-	0.411	-0.250
PZT	0.688	-	-	-	<0.01	0.635	-	-	0.251	0.365
SPS	0.808	0.468	-	0.634	-	-0.040	0.650	-	0.586	0.412
TPS	1.000	1.00	-	0.714	-	0.281	0.700	0.38	0.485	0.767
PS	0.714	-	-	0.610	-	0.437	-	-	-	-
ZS	0.737	0.857	0.39	0.600	0.06	0.374	0.541	0.11	0.166	0.357

* experienced observer

**inexperienced observer