uc3m | Universidad **Carlos III** de Madrid

Master Degree in Statistics for Data Science
Academic Year 2021-2022

Master Thesis

“Implementation of an out-of-sample classification rule in a smartphone App for Severe Acute Malnutrition (SAM) diagnosis”

Álvaro Pérez Romero

Ana Arribas Gil
Laura Medialdea Marcos
September 2022, Madrid

Abstract

Every year, millions of children die because of undernutrition all over the world. This population group is specially vulnerable to this condition given that, at the early stages of life, the human body is yet to be developed. Therefore, it is necessary to propose and implement solutions that might mitigate this problem.

In this regard, Action Against Hunger started developing on 2015 an easy-to-use smartphone-based app called SAM Photo Diagnosis App® that is able to detect the nutritional status of children under five years old by simply analyzing their body shape. To achieve this, SAM photo app uses Geometric Morphometric (GM) techniques, which are a collection of tools widely used for visualization and quantification of shape changes among biological. Therefore, given an initial sample of children whose nutritional status had been registered and their body shape information summarized in a list of landmarks, GM techniques such as Procrustes analysis (and possible size effect removal) are applied in order to standardize these landmarks (by performing alignment of the landmarks within all the children). In the new set of coordinates obtained, known as set of Procrustes coordinates, the application of a supervised classification rule is straightforward. Nonetheless, the application of the same classification rule to an out-of-sample observation might be problematic taking into account that its landmarks are yet to be standardized. This said, given that there is no consensus in the literature about which algorithm is the most suitable for performing out of sample classifications, since this problem has not yet been addressed in depth, this Master Thesis aims to introduce 2 standardization techniques that will allow us to align any out-of-sample observation with the rest of Procrustes observations. These 2 techniques are the following: registration of an out-of-sample individual to the mean shape and registration of an out-of-sample individual to the median shape.

Key words: SAM, supervised learning, undernutrition, Geometric Morphometrics, Procrustes analysis.

Contents

Chapter 1

Introduction

According to the World Health Organization (WHO, 2020), **malnutrition** refers to "deficiencies, excesses or imbalances in a person's intake of energy and/or nutrients". Furthermore, there exist 2 different types of malnutrition: "One is 'undernutrition'—which includes stunting (low height for age), wasting (low weight for height), underweight (low weight for age) and micronutrient deficiencies or insufficiencies (a lack of important vitamins and minerals). The other is overweight, obesity and diet-related noncommunicable diseases (such as heart disease, stroke, diabetes, and cancer)".

Taking into account that the nutritional status (Grellety, Golden, 2016; Myatt, Duffield, Seal, Pasteur, 2009) is a factor related to the development in the human body, then it must be paid extreme attention to the diet every child takes so that they are assured a proper growth process. However, according to the World Health Organization (WHO, 2020) "in 2020, 149 million children under 5 were estimated to be stunted, 45 million were estimated to be wasted and 38.9 million were overweight or obese". And, even worse, "around 45% of deaths among children under 5 years are linked to undernutrition". In this scenario, it is clear that many improvements need to be done in order to fight malnutrition worldwide.

This being said, in this Master's Thesis we will focus on the wasting sub-form of undernutrition in children under five years old. This sort of malnutrition occurs when a person "has not had food of adequate quality and quantity and/or they have had frequent or prolonged illnesses". Besides, "wasting in children is associated with a higher risk of death if not treated properly". As a result, it is crucial to spot sufficiently in advance any child affected by this condition so that the child receives the proper health care. In order to diagnose wasting undernutrition, 3 different anthropometric techniques are usually applied (de Onis Habicht, 1996; WHO, 2013):

1) Detecting oedema by pressing with a finger the child's feet during 10 seconds and checking whether the pressed surfaces retracts or not. When oedema is found, the child needs immediate health care.

2) Calculating the weight-for-height z-score (WHZ) taking into account, for a given length (cm), the sample mean and sample standard deviation in weight (kg) from a reference sample (generally supplied by WHO). The following nutritional status are obtained:

- Normal nutrition (NOR): WHZ ≥ -1.
- Risk of undernutrition (RIS): $-2 \leq$WHZ< -1.
- Moderate wasting/acute malnutrition (MAM): $-3 \leq$ WHZ < -2.
- Severe wasting/acute malnutrition (SAM): WHZ< 3.

3) Measuring the mid-upper (left) arm circumference (mm) mid-way between the tip of the elbow and the tip of the shoulder. To ease this task, there exist paper bands typically coded into 3 or 4 colors (green, yellow, orange, red) and depending on which color the child's arm falls within, then different nutritional status are obtained:

- Green - Normal nutrition (NOR): MUAC $\geq$ 135 mm.
- Yellow - Risk of undernutrition (RIS): 125 mm $\leq$ MUAC $<$ 135 mm.
- Orange - Moderate wasting/acute malnutrition (MAM): 115 mm $\leq$ MUAC $<$ 125 mm.
- Red - Severe wasting/acute malnutrition (SAM): 115 mm $<$MUAC.

A representative example of how this paper bands work is shown in the next figure:

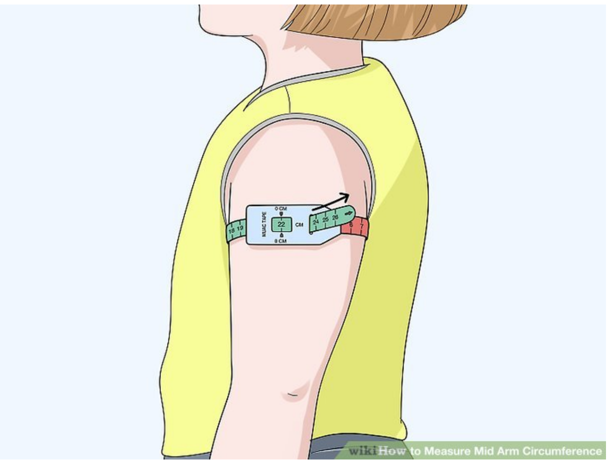


Figure 1.1: (WikiHow, 2022) Child belonging to NOR nutritional status. The arm falls within green color.

These last 2 techniques have some drawbacks associated (Medialdea et al, 2019). From one hand, examinations are time-consuming and they rely on specific instruments that need to be calibrated properly (such as stadiometers, measuring tapes, calipers or weighing scales). And, from the other hand, collaboration from the patient (by remaining still) as well as professional supervision is needed in order to avoid measurement errors and, hence, inaccurate results. It is very important to bear in mind that any missclassification made regarding to the nutritional status might be fatal for many (undernourished) children who do not receive a proper health care because of it. In spite of this, health agents working in communities from low- or middle-income countries are sometimes not adequately trained. Therefore, it is necessary to provide communities easy-to-use alternative methods to diagnose undernutrition in children.

Under this pretext, Action Against Hunger started developing on 2015 an Smart Phone-based app that automatically detects the level of wasting undernutrition (whether NOR, RIS, MAM or SAM) in children between 6 and 59 months by simply taking a photograph of their left arm and analyzing its shape. This app is called **SAM Photo Diagnosis App®** and it is, to date, still being developed (stage 3) (Action Against Hunger, n.d.). Furthermore, it is meant to work without need of internet connection (being this feature key in order to diagnose undernutrition even in underdeveloped places where connectivity barely exists).

The starting hypothesis for developing this application was that that there exist differences in body shape between children of different nutritional status, sex and age. Therefore, given that Geometric Morphometric (GM) techniques are a collection of tools widely used for visualization and quantification of shape changes among biological forms and that these forms can be represented by bi- or three-dimensional cartesian points or landmarks that summarize their shape information (Adams, Rohlf, Slice, 2004), then it was decided to use these techniques to analyze the shape in the body of children between 6 and 59 months. In order to do that, many photographs were taken from different views of their bodies and then they were translated into different sets of bi-dimensional landmarks. However, among all the possible views considered, the supine position resulted to be the view that better summarized the information about the fat and muscles disposition (Medialdea et al, 2019). Besides, within the supine position, the 4 limbs and the trunk were analyzed separately, resulting the upper left limb the part of the body more informative regarding to the existence of severe acute malnutrition (Medialdea et al, 2021).

Taking into account that GM techniques, in combination with left arm landmarks, facilitate the study of undernutrition in children under five, then it was decided to use both 'ingredients' in order to develop SAM photo app classification method. However, the creation of a classifier in GM is quite challenging considering that, until now, GM techniques have been used for in-sample shape analysis but not often for out-of-sample classification purposes. This said, the following strategy has been designed in order to make SAM photo app work:

To start with, there needs to be a training sample of children under five years old whose nutritional status had been previously obtained (using traditional methods such as WHZ or MUAC). Then, for each child, a photograph of the left arm needs to be taken so that SAM photo app can translate it into a configuration of landmarks which summarizes its shape. Now, given that photographs might have been taken from different heights, positions and angles, then landmark configurations need to be re-scaled, translated and rotated (as a whole) so that all the arms are aligned within children. In order to do that, Procrustes analysis techniques are applied and a new set of coordinates, called Procrustes coordinates, is obtained. Right after this 'standardization' had been performed (and, maybe, size effect removal), the training dataset is ready to be supplied to any supervised learning classification model. Nonetheless, the models fitted in the Procrustes coordinates are not suitable to be tested in an out-of-sample/test child taking into account that the out-of-sample raw coordinates obtained by SAM photo app (after the associated photograph had been translated) might be far away from the Procrustes coordinates. Therefore, Procrustes analysis techniques are applied again and these out-of-sample coordinates are projected to the training dataset of Procrustes coordinates by registering the out-of-sample configuration to the (Procrustes sample) mean shape or to the (Procrustes sample) median shape. Finally, the classification rule is applied and the nutritional status of the out-of-sample child is obtained.

Chapter 2

Classification in Geometric Morphometrics

In this chapter, we will briefly introduce the Geometric Morphometric techniques needed to train a supervised classification model in a set of shapes.

Typically, the way we distinguish one object from another is by looking at the shape they have. Sometimes differences in shape can clearly be noticed (e.g., when comparing a triangle versus a square or a cat versus a dog) but some other times it is not straightforward. For instance, if we take a look at the figure below, can we really say which upper limb belongs to an undernourished child and which one to a healthy child?

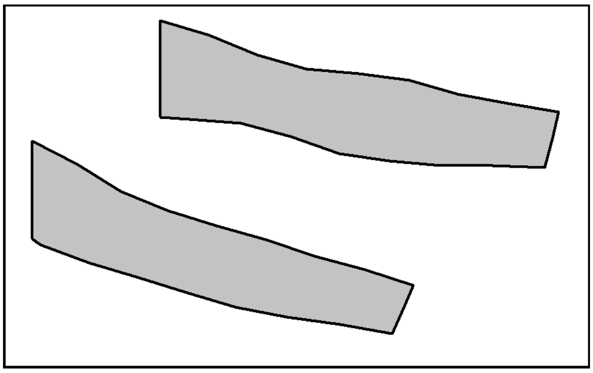


Figure 2.1: Upper left limb's representation of a child belonging to group SAM (bottom) and that of a child belonging to group NOR (top). Notice that the actual representation of the arm would be the photograph itself, but it is not shown for privacy reasons.

To answer questions like the one stated above, Geometric Morphometric (GM) techniques arise. To begin with, let us introduce the concept of shape (Kendall, 1977):

Definition 2.0.1. *The* ***shape*** *of an object is all the geometrical information that is invariant under transformations of scale, location and rotation.*

Therefore, 2 objects have the same shape whenever any of them can be scaled, translated and rotated so that it matches exactly the other one (e.g., any 2 equilateral triangles in the euclidean plane). Nonetheless, in our case it is almost impossible to find 2 children who might have an arm with the exact same shape. There are two basic reasons for this: firstly, there always exist differences in the body shape and, secondly, there might exist measurement errors in the dataset. Owing to this, we should focus on detecting which upper limbs have a **similar shape** instead of which limbs have the exact same shape. For that purpose, we need to compare arms' shapes using some sort of additional information apart from a mere graphical representation. More precisely, we are going to locate on each arm a list of bidimensional points that represent the geometrical information contained in them. These particular points are called landmarks (Bookstein, 1991):

Definition 2.0.2. *A* ***landmark*** *is a point of correspondence located on a biological form that matches between and within populations.*

Definition 2.0.3. *A* ***configuration*** *is a set of landmarks summarizing the shape of a form. A* ***configuration matrix*** *$X \in M_{k\times m}(\mathbb{R})$ (the vector space of real matrices) is a real matrix containing the shape information of a form in k landmarks (rows) of dimension m. The shape that X represents is denoted by $[X]$.*

Note 2.0.1. *Usually, an initial list of landmarks having some sort of scientific or mathematical meaning are located on geometrical forms, and, after that, a second list of landmarks called semi-landmarks (Bookstein, 1997) are located, relative to the initial landmarks' positions, on areas which are hard to capture (like smooth curves and surfaces) with the purpose of completing the geometrical information given by the initial landmarks. Once all the landmarks have been located, the configuration is ready.*

Following the example given in Figure 2.1, we can locate a set of landmarks on the given arms in order to represent their shape. Let us see the next Figure:

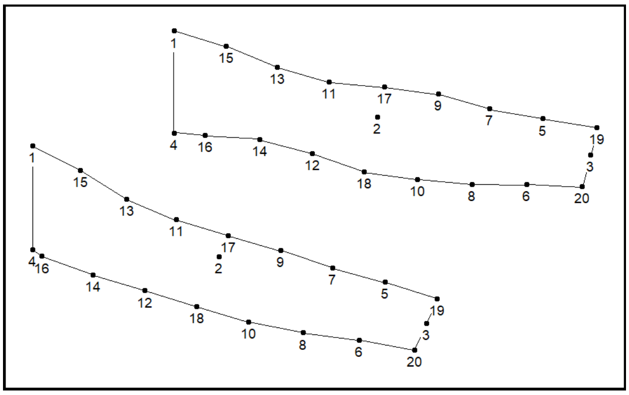


Figure 2.2: Landmarks representation of the arms given in Figure 2.1. Additionally (and for illustration purposes), a line between each pair of consecutive landmarks belonging to the outline has been plotted.

We have already commented in the introduction that SAM photo app automatically translates a photograph of a left arm into a set of landmarks. Furthermore, SAM photo app is designed to suggest an ideal position from where the photographs of the upper left limbs need to be taken. Still, differences in the scale, angle and position of the coordinates exist from photograph to photograph. Therefore, we need to correct them before we can even fit a classification model.

2.1 Size, rotation and translation

In this section we will illustrate how the correction of size, rotation and location can be achieved among arms by manually applying the necessary transformations.

This being said, depending on the height from where a photograph is taken and also depending on the real size of a child's arm, the scale in the coordinates obtained by SAM photo app might vary.
For instance, if we take a look at the 2 arms given in the figure below and compare them, then it is clear that the ranges of the X values and the Y values are different, i.e., the limbs' sizes are different:

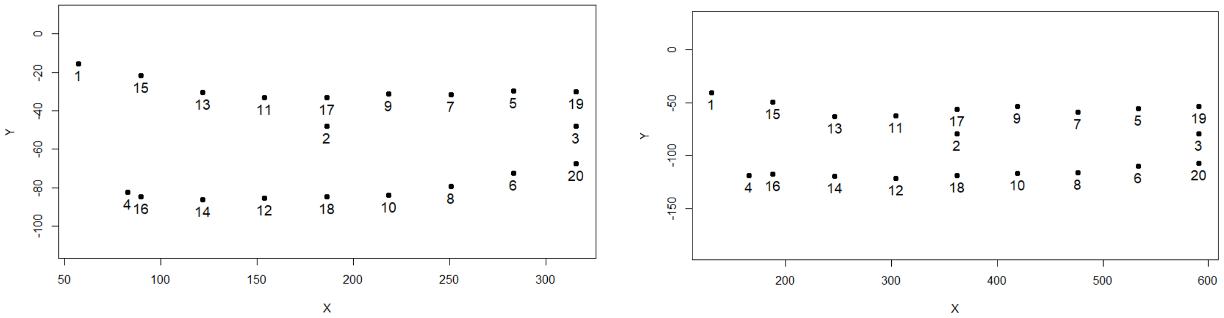


Figure 2.3: Landmarks' coordinates (obtained by SAM Photo Diagnosis App®) of 2 different arms.

A possible solution to fix this issue is to standardize all the configuration matrices to the same size (the reader must bear in mind that this way of proceeding is naive and it is only intended for illustration purposes throughout this section; alternative methods will be presented in the next section). However, before we can even do that, we need to introduce the following definition (Bookstein, 1986):

Definition 2.1.1. *Let $X \in M_{k\times m}(\mathbb{R})$ be a configuration matrix. A **size measure** g of X is any positive real valued function $g : X \to \mathbb{R}$ such that, for all $a > 0$:*

$$g(aX) = ag(X).$$

There exist many possible functions that might be used as size measure (for instance, the baseline size or the area of the convex hull; Dryden, Mardia, 2016). However, in this Thesis we have decided to choose the centroid size as our preferred size measure as it is widely used in GM (Kendall 1984; Bookstein 1986;...):

Definition 2.1.2. *Let $X \in M_{k\times m}(\mathbb{R})$ be a configuration matrix. The **centroid size** s of X is given by:*

$$s(X) := \sqrt{\sum_{i=1}^{k} \|X_{i,*} - \overline{X}\|^2},$$

*where $X_{i,*}$ is the the i-th row of X, i.e., the i-th landmark of X, and $\overline{X} = (\overline{X_{*,1}}, ..., \overline{X_{*,m}})$ is the vector of column means, this is, the **centroid**.*

Note 2.1.1. *It holds that $s(aX) = as(X)$ for all $a > 0$.*

Therefore, for a given configuration matrix, we can divide by its centroid size in order to 'standardize' to size 1. Let us see an example in the figure below:

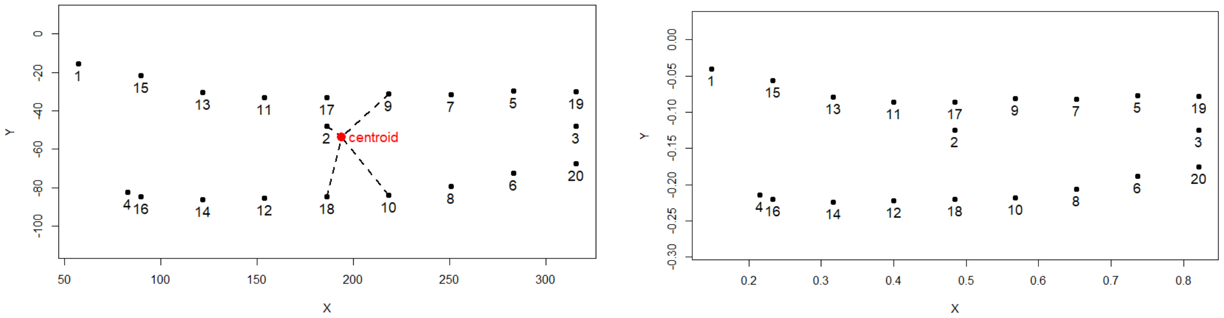


Figure 2.4: Landmarks' coordinates before (left image; size equal to 384.3) and after (right image; size equal to 1) standardization of the size has been carried out for a given arm. Notice that the shape remains the same.

Once the issue regarding to the scale has been addressed and all the configuration matrices are somehow equally sized, the next thing that needs to be fixed is the location of the landmarks. Hence, we need to perform **translation** in the arms's coordinates so that the configuration matrices are reasonably close to each other.

An easy way to translate a configuration matrix $X \in M_{k\times m}(\mathbb{R})$ is to add to each row of X a constant vector $\gamma = (\gamma_1, ..., \gamma_m)^T \in \mathbb{R}^m$. However, if centering about the origin $\mathbf{0} \in \mathbb{R}^m$ is rather desired, then it is necessary to transform X as follows (Helmert, 1876):

$$X_C = (I_k - \frac{1}{k}1_k 1_k^T)X,$$

where I_k is the identity matrix of order k and $1_k^T = (1, \overset{k)}{\ldots}, 1) \in \mathbb{R}^k$.

As a result, we can apply this transformation in all the configuration matrices available in order to fix the location problem. Let us see the following example:

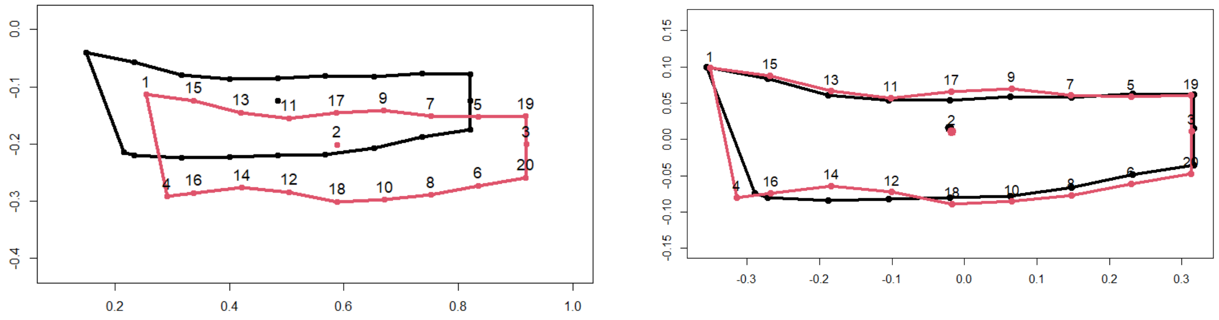


Figure 2.5: Landmarks' coordinates of the upper left limbs given in Figure 2.3 before (left image) and after (right image) translation about the origin has been performed.

After size removal and translation steps have been carried out, the last thing we need to perform is **rotation** (bear in mind that photographs might have been taken from different angles). More precisely, we need that, among the arms, the landmarks labelled as 1 are all close to each other, that the landmarks labelled as 2 are all close to each other, and so on. In order to achieve this, we need to post-multiply each configuration matrix $X \in M_{k\times m}(\mathbb{R})$ by the appropriate rotation matrix $\Gamma \in M_{m\times m}(\mathbb{R})$, i.e., $X\Gamma$.

Definition 2.1.3. *A **rotation matrix** Γ is an orthogonal matrix (i.e., an square matrix such that $\Gamma^T\Gamma = \Gamma\Gamma^T = I_m$) satisfying that $det(\Gamma) = +1$. When $m = 2$, the rotation matrix can be parameterized as*

$$\Gamma(\theta) = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix},$$

where $-\pi \leq \theta < \pi$.

Note 2.1.2. *Given a configuration matrix $X \in M_{k\times m}(\mathbb{R})$ such that it is centered about the origin, then $X\Gamma$ is the anticlockwise rotation of X θ radians about the origin.*

In Figure 2.5 (right sub-figure) it is not clear whether we should perform rotation in the given configuration matrices because each pair of landmarks seems to be very close. Conversely, if we take a look at the next figure, then it seems that rotation needs to be performed in order to achieve a better alignment:

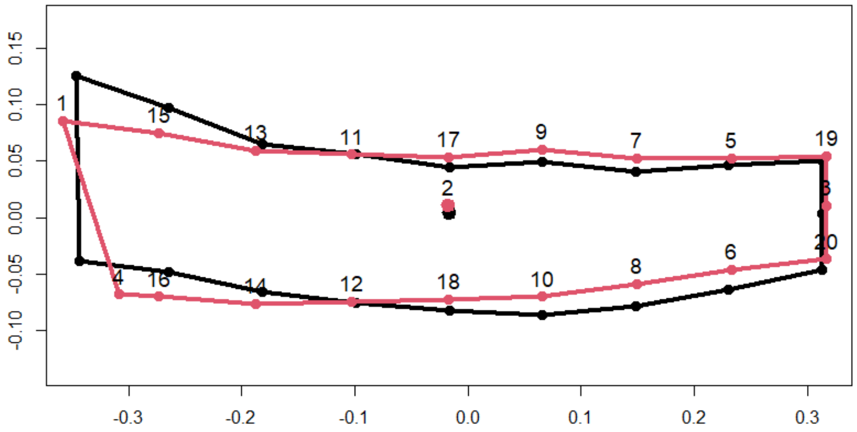


Figure 2.6: Landmarks' coordinates of 2 different arms. Better adjustment might be obtained if we rotated just a little bit to the right (and with respect to the origin) the red form

Summing up, in this section we have seen that it is possible to align any list of configuration matrices by performing a few transformations to each of them manually.

However, this is a simple and tedious approach and, therefore, different sort of techniques need to be proposed. This said, let us check the next section.

2.2 Generalized Procrustes analysis

In this section, we will address the correction of the scale, location and rotation among configuration matrices in a more suitable way. More precisely, we will use Procrustes Analysis techniques.

The term Procrustes analysis was first used in 1962 (Hurley, Cattell, 1962) and it is a method of shape alignment that owes its name to a bandit from the Greek mythology called Procrustes. This bandit offered travellers a bed for staying the night and then, depending on the height of his victims, he fit them to the size of the bed by either stretching their limbs or cutting them off. Following this analogy, we are going to transform the children's arms so that their configurations of landmarks match as much as possible.

There exist many variations in Procrustes Analysis techniques. However, in this Master's Thesis we will focus only on 2 of them: full Ordinary Procrustes Analysis and full Generalized Procrustes Analysis. This said, let us begin with the first one.

Suppose that we have a pair of configuration matrices supplied by SAM photo app and that we want to align or superimpose one of those configuration matrices to the another (the latter remaining still, as opposed to the previous section, where every configuration matrix was translated about the origin). Then, we would need to estimate the optimal scaling, translation and rotation parameters (similarity parameters) that allow us to achieve this:

Definition 2.2.1. *Let $X_1, X_2 \in M_{k\times m}(\mathbb{R})$ be 2 configuration matrices. The method of full Ordinary Procrustes Analysis or **full OPA** performs least squares in order to estimate the similarity parameters γ, Γ and β that minimize the distance*

$$D^2_{OPA}(X_1, X_2) := \|X_2 - \beta X_1 \Gamma - 1_k \gamma^T\|^2, \tag{2.1}$$

*where $\|.\|$ is the Euclidean norm (i.e., $\|A\| = (\sum_{i=1}^{k}\sum_{j=1}^{m} a_{i,j}^2)^{\frac{1}{2}}$ for $A \in M_{k\times m}(\mathbb{R})$), $1_k = (1, \overset{k)}{\ldots}, 1)^T \in \mathbb{R}^k$, $\beta > 0$ is a scale parameter, Γ is an $m \times m$ rotation matrix and γ is an $m \times 1$ location vector. The minimum of this equation, $OSS(X_1, X_2)$, stands for **ordinary (Procrustes) sum of squares**.*

Definition 2.2.2. *The **full OPA registration** of X_1 to X_2 is given by:*

$$X_1^R := \hat{\beta} X_1 \hat{\Gamma} + 1_k \hat{\gamma}^T,$$

where $\hat{\beta}, \hat{\Gamma}, \hat{\gamma}$ are the minimizing parameters in (2.1).

Note 2.2.1. *For simplification purposes we assume, without loss of generality, that all the configuration matrices are centered about the origin $\mathbf{0} \in \mathbb{R}^m$.*

Note 2.2.2. *The previous optimization problem has a closed-form solution (Dryden, Mardia, 2016) given by:*

$$\begin{aligned}
&\hat{\gamma} = 0,\\
&\hat{\Gamma} = UV^T,\\
&\hat{\beta} = trace(X_2^T X_1 \hat{\Gamma})/trace(X_1^T X_1),
\end{aligned}$$

where U, V are the matrices obtained from the singular value decomposition of $X_2^T X_1$ after scale and translation have been removed (the latter being already removed) from X_1 and X_2. On the other hand, we need to take into account that, depending on which configuration *matrix we decide to register first, whether X_1 to X_2, or X_2 to X_1, then the solution might vary.*

So, by simply performing least squares techniques, we can register one configuration matrix to another. In order to understand this, let us check the next figure:

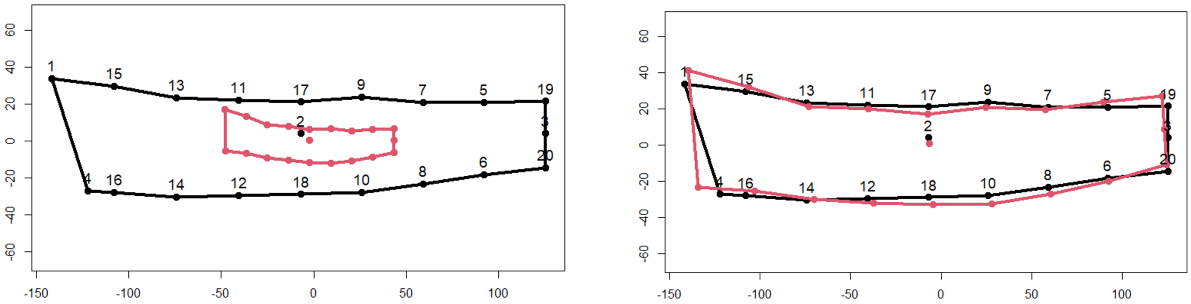


Figure 2.7: Landmarks' coordinates of an arm (red) before (left image) and after (right image) obtaining the registration to another arm (black) by full OPA method. The estimated parameters are $\hat{\beta} = 2.85$, $\hat{\gamma} = 0$ and $\hat{\theta} = 3.43$ degrees.

Although full OPA is a powerful tool that allows us to align any pair of configuration matrices (being this feature key in order to develop the next chapter), this technique is not enough if alignment over a list of more than 2 configuration matrices is rather desired. And, ideally, SAM photo app would need to be supplied a dataset of limbs properly aligned before models can even be trained. This said, the method that substitutes full OPA and that will allow us to standardize our arms' configuration matrices is called full Generalized Procrustes Analysis (Gower 1975; Rohlf, Slice 1990):

Definition 2.2.3. *Let $X_1, ..., X_n \in M_{k\times m}(\mathbb{R})$ be configuration matrices. The method of full Generalized Procrustes Analysis or* ***full GPA*** *performs least squares in order to estimate the parameters γ_h, Γ_h, β_h, $h = 1, ..., n$, and $\hat{\mu}$ that minimize the total sum of squares*

$$G(X_1, ..., X_n) = \sum_{h=1}^{n} \|(\beta_h X_h \Gamma_h + 1_k \gamma_h^T) - \hat{\mu}\|^2 \tag{2.2}$$

subject to a constraint in the sizes. There are many possible constraints, but the following one is very useful in order to prevent the estimated $\hat{\beta}_h$ becoming close to 0:

$$\sum_{h=1}^{n} s^2(\beta_h X_h \Gamma_h + 1_k \gamma_h^T) = \sum_{h=1}^{n} s^2(X_h),$$

where s is the centroid size.

Definition 2.2.4. *The* ***full Procrustes coordinates*** *of X_h are given by:*

$$X_h^P := \hat{\beta}_h X_h \hat{\Gamma}_h + 1_k \hat{\gamma}_h^T, \quad h = 1, ..., n,$$

where $\hat{\beta}_h > 0$, $\hat{\Gamma}_h$, $\hat{\gamma}_h^T$, are the minimizing parameters in (2.2).

Note 2.2.3. *Notice that, in full GPA, we are estimating as many sets of parameters as configurations available. Furthermore, we need to take into account that, given that X_h is centered for all h, then $\hat{\gamma}_h = 0$, and we only need to estimate $\hat{\mu}$, $\hat{\beta}_h$, $\hat{\Gamma}_h$. In order to do that, we need to perform iterative methods (except for $m = 2$, where an explicit eigenvector solution exists). An example algorithm is given below (Dryden, Mardia, 2016):*

1. *Initialize $X_h^P = X_h$, for all $h = 1, ..., n$.*
2. *Calculate $\hat{\mu} = \frac{1}{n}\sum_{h=1}^{n} X_h^P$.*
3. *Update $X_h^P = (X_h^P)^R$ $(h = 1, ..., n)$, where $(X_h^P)^R$ is the registration of X_h^P to $\hat{\mu}$.*
4. *Repeat steps 2 and 3 until $\sum_{h=1}^{n} \|X_h^P - \hat{\mu}\|^2$ is below a tolerance parameter.*

Note 2.2.4. *It is very important to bear in mind that the parameter of primary interest estimated in full GPA is precisely $\hat{\mu}$. And, what does this parameter represents? Well, let us suppose that the sample of configuration matrices $X_1, ..., X_n$ has been generated from a random configuration matrix with population mean $\mu \in M_{k\times m}(\mathbb{R})$. Then,* ***whenever we perform full GPA, we are actually estimating the shape of the population mean,*** $[\mu]$***, with an 'average' shape obtained from the sample,*** $[\hat{\mu}]$. *This said, $\hat{\mu}$ has the same shape as $\frac{1}{n}\sum_{h=1}^{n} X_h^P$ (Dryden, Mardia, 2016), i.e., the arithmetic mean of the Procrustes configuration matrices.*

Note 2.2.5. *Although a given shape might be represented by uncountable infinite possible configuration matrices, from now on, and for simplification purposes, we will represent $\hat{\mu}$ as*

$$\hat{\mu} := \frac{1}{n}\sum_{h=1}^{n} X_h^P,$$

where $\hat{\mu}_{i,j} = \frac{1}{n}\sum_{h=1}^{n} (X_h^P)_{i,j}$ for all i, j.

In a few words, full GPA is a GM technique that not only estimates the mean shape from a sample but also it aligns all the sample observations with that average shape.

As a result, we will be able to properly align all the in-sample children arms (and, hence, to fit a supervised learning model) and also to understand how does an average child's arm looks like in our dataset. This will help us to visualize how, for different nutritional status, the shapes differ from this estimated mean shape and, hence, to obtain conclusions based on the shape differences observed.

In the figure below, a demonstration example of how full GPA actually works is shown:

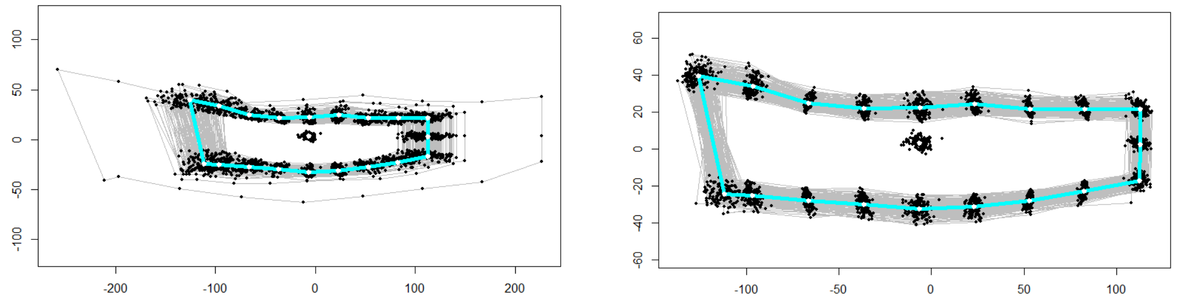


Figure 2.8: Landmarks' coordinates of a given group of children before (left image) and after (right image) full GPA has been performed. The sample mean estimated is given in blue color.

2.3 Classification

At this point, we are ready to define a classification rule in the set of Procrustes coordinates obtained. In order to do that, we just need to think of a list of Procrustes configuration matrices $X_1^P, ..., X_n^P$ as a matrix or dataset $P \in M_{n\times k\cdot m}(\mathbb{R})$, where the h-th row of P corresponds with the (flattened) h-th Procrustes configuration matrix, X_h^P, and where each consecutive set of m columns represent a landmark (in the sense that, for $m = 2$ and h fixed, $P_{h,1}$ and $P_{h,2}$ are the x and y coordinates of the first landmark of X_h^P, i.e., $(X_h^P)_{1,1}$ and $(X_h^P)_{1,2}$, that $P_{h,3}$ and $P_{h,4}$ are the x and y coordinates of the second landmark of X_h^P, i.e., $(X_h^P)_{2,1}$ and $(X_h^P)_{2,2}$, and so on).

On the other hand, notice that, although full GPA corrects location, scale and rotation in the left arm landmarks obtained by SAM photo app, it does not correct the possible effect that the actual size of an arm (in the reality) might have on its shape, i.e., the effect associated to Allometry (Dryden, Mardia, 2016). More precisely, we need to take into account that, given that children's age varies from 5 to 59 months, so does their body size and, hence, their body proportions. As a result, different arm shapes might be obtained (even for children belonging to the same nutritional status) owing to the growth process of children. For this reason, it might also be a good idea to correct the effect associated to Allometry in order to improve the performance of our classifiers.

To achieve this, we need to substitute the dataset P for a new dataset, $\tilde{P}$, which rows contain the residual Procrustes coordinates after the following regressions have been performed:

Definition 2.3.1. *Let $X_1^P, ..., X_n^P$, and $s_1, ..., s_n \in \mathbb{R}$, be, respectively, the Procrustes coordinates and the centroid sizes of $X_1, ..., X_n \in M_{k\times m}(\mathbb{R})$. Let us also fit, for each $i = 1, ..., k$, $j = 1, ..., m$, the following regression model in the Procrustes coordinates:*

$$Y_{i,j}^P = a_{i,j} + b_{i,j} \cdot s(Y),$$

where $s(Y)$ is the centroid size of the random matrix $Y \in M_{k\times m}(\mathbb{R})$ and $Y_{i,j}^P$ is the (i, j) - coordinate of Y (which is a random variable). Then, for $h = 1, ..., n$, the h-th row of $\tilde{P}$ is given by the (flattened) residual configuration matrix $X_h^P - \widehat{X_h^P}$, where

$$(\widehat{X_h^P})_{i,j} = \widehat{a_{i,j}} + \widehat{b_{i,j}} \cdot S(X_h), \quad \textit{for each } i, j.$$

Note 2.3.1. *Bear in mind that, if Allometry is performed in order to obtain the final training dataset, then, for each out-of-sample (already aligned) observation Z, the coordinates to be supplied to the classifier would be Z-$\hat{Z}$, where $\hat{Z}$ is the estimation of Z using the Allometry regression models.*

Chapter 3

Out-of-sample projection

In the previous chapter we have seen that full GPA allows us to standardize or align a set of raw arms' landmarks coordinates $X_1, ..., X_n$ and to obtain a new set of them called Procrustes coordinates $X_1^P, ..., X_n^P$ which are ready to be used in order to define a classification rule. In this regard, supervised classifications models such as Linear Discriminant Analysis might be trained in order to analyze the in-sample shapes (Sandoval Oyanedel, Diaz, Manríquez, 2021). However, it is not evident whether these classification rules will properly classify an out-of-sample arm or observation that might arise taking into account that the raw out-of-sample coordinates obtained by SAM photo app might be far from being aligned with the dataset of Procrustes coordinates. In order to understand this, we can take a look at the next figure:

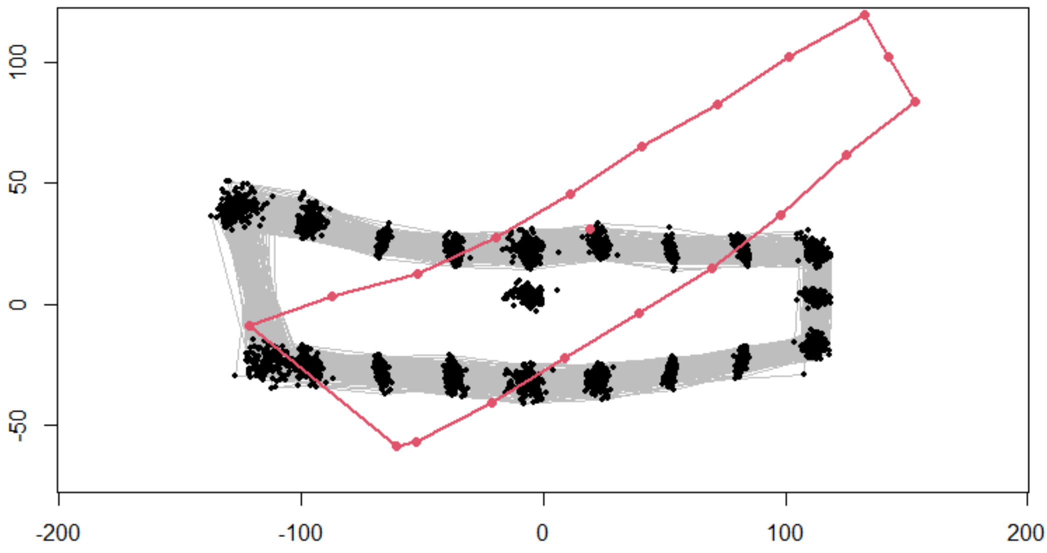


Figure 3.1: Procrustes coordinates of the arms belonging to Figure 2.8 (right sub-figure) along with an out-of-sample arm (red).

This said, a naive approach to align an out-of-sample observation X_{n+1} with $X_1^P, ..., X_n^P$ would be to perform full GPA again but, now, considering the dataset of in-sample raw coordinates along with the out-of-sample raw coordinates, i.e., $X_1, ..., X_n, X_{n+1}$. However, taking into account that this minimizing problem is different to the initial one (as the average sample shape of $X_1, ..., X_n, X_{n+1}$ is influenced by this out-of-sample observation), then the new in-Procrustes sample coordinates obtained, $X_1'^P, ..., X_n'^P$, would be different to the initial ones, $X_1^P, ..., X_n^P$. And, as a result, the already fitted models in $X_1^P, ..., X_n^P$ would not longer share the same set of training Procrustes coordinates. Owing to this, and also owing to the fact that this approach might be computationally expensive (as many minimizing problems need to be solved as existing out-of-sample observations), then it is necessary to propose alternative methods. Instead, a better approach would be to simply align X_{n+1} with the population mean μ of the random configuration matrix X that has generated the sample $X_1^P, ..., X_n^P$ (not $X_1, ..., X_n$). However, given that this is just the ideal world, the most natural way to proceed is to simply align X_{n+1} with the sample average $\hat{\mu}$ of $X_1^P, ..., X_n^P$. On the other hand, it might happen that the sample is heterogeneous and that a more robust estimation is rather desired. In this case, we might also align X_{n+1} with the (sample) median of $X_1^P, ..., X_n^P$. This said, in the following sections we will address these 2 methods:

3.1 Registration to the mean shape

Definition 3.1.1. *Let $X_1^P, ..., X_n^P$ the Procrustes coordinates of $X_1, ..., X_n \in M_{k\times m}(\mathbb{R})$ and let $X_{n+1} \in M_{k\times m}(\mathbb{R})$ be an out of sample observation. Then, the* ***registration of*** X_{n+1} ***to the (Procrustes sample) mean shape*** *is given by*

$$X_{n+1}^R = \hat{\beta} X_{n+1} \hat{\Gamma} + 1_k \hat{\gamma}^T,$$

where $\hat{\beta}, \hat{\Gamma}, \hat{\gamma}$ are the minimizing similarity parameters estimated in equation (2.1) *after registering X_{n+1} to $\hat{\mu}$ with full OPA (recall that $\hat{\mu}$ is the arithmetic mean of $X_1^P, ..., X_n^P$).*

So by calculating the sample mean of the training set of Procrustes configuration matrices and then performing full OPA in order to obtain the registration of an out-of-sample configuration matrix to the sample mean, then we obtain a projection of the out-of-sample individual to the training Procrustes coordinates.

Furthermore, this registration method is very complete because **it is equivalent to obtaining the configuration matrix closest to all the Procrustes configuration matrices**. In the sense that X_{n+1}^R minimizes the sum of squared distances d_{OPA}^2 to all X_h^P. Let us see the following result:

Proposition 3.1.1. *Let $X_1^P, ..., X_n^P$ be the Procrustes coordinates of $X_1, ..., X_n \in M_{k\times m}(\mathbb{R})$ and $X_{n+1} \in M_{k\times m}(\mathbb{R})$ be an out-of-sample observation. Then, obtaining the registration of X_{n+1} that minimizes the sum of squared distances d_{OPA}^2 to all the Procrustes coordinates $X_1^P, ..., X_n^P$ is equivalent to obtaining the registration of X_{n+1} to $\hat{\mu}$ by full OPA.*

Proof. In order to prove it, we need to check the following (where β, Γ, γ are the similarity parameters):

$$\begin{aligned}&\arg\min_{\beta,\Gamma,\gamma} \sum_{h=1}^{n} d_{OPA}^2(X_{n+1}, X_h^P) = \arg\min_{\beta,\Gamma,\gamma} \sum_{h=1}^{n} \|X_h^P - \beta X_{n+1}\Gamma - 1_k\gamma^T\|^2 = \\ &\arg\min_{\beta,\Gamma,\gamma}\ d_{OPA}^2(X_{n+1}, \hat{\mu}) = \arg\min_{\beta,\Gamma,\gamma}\ \|\hat{\mu} - \beta X_{n+1}\Gamma - 1_k\gamma^T\|^2\end{aligned}$$

This said, for a given set of similarity parameters β, Γ, γ, we have the following equations:

$$\begin{aligned}
\sum_{h=1}^{n} d^2_{OPA}(X_{n+1}, X_h^P) &= \sum_{h=1}^{n}\sum_{i=1}^{k}\sum_{j=1}^{m}((X_h^P)_{i,j} - \beta(X_{n+1})_{i,*} \cdot \Gamma_{*,j} - 1_k\gamma^T)^2 \\
&= \sum_{h=1}^{n}\sum_{i=1}^{k}\sum_{j=1}^{m}((X_h^P)_{i,j} - \hat{\mu}_{i,j} + \hat{\mu}_{i,j} - \beta(X_{n+1})_{i,*} \cdot \Gamma_{*,j} - 1_k\gamma^T)^2 \\
&= \sum_{h=1}^{n}\sum_{i=1}^{k}\sum_{j=1}^{m}((X_h^P)_{i,j} - \hat{\mu}_{i,j})^2 + \\
&+ \sum_{h=1}^{n}\sum_{i=1}^{k}\sum_{j=1}^{m}(\hat{\mu}_{i,j} - \beta(X_{n+1})_{i,*} \cdot \Gamma_{*,j} - 1_k\gamma^T)^2 + \\
&+ 2 \cdot \sum_{h=1}^{n}\sum_{i=1}^{k}\sum_{j=1}^{m}((X_h^P)_{i,j} - \hat{\mu}_{i,j}) \cdot (\hat{\mu}_{i,j} - \beta(X_{n+1})_{i,*} \cdot \Gamma_{*,j} - 1_k\gamma^T)
\end{aligned}$$

Now, within the last expression, the first addend does not depend on any of the similarity parameters and $\hat{\mu}$ is fixed. Furthermore, for the third addend it holds the following:

$$\begin{aligned}
& 2 \cdot \sum_{h=1}^{n}\sum_{i=1}^{k}\sum_{j=1}^{m}((X_h^P)_{i,j} - \hat{\mu}_{i,j}) \cdot (\hat{\mu}_{i,j} - \beta(X_{n+1})_{i,*} \cdot \Gamma_{*,j} - 1_k\gamma^T) \\
= \; & 2\sum_{i=1}^{k}\sum_{j=1}^{m}(\hat{\mu}_{i,j} - \beta(X_{n+1})_{i,*} \cdot \Gamma_{*,j} - 1_k\gamma^T) \cdot \sum_{h=1}^{n}((X_h^P)_{i,j} - \hat{\mu}_{i,j}) \\
= \; & 0
\end{aligned}$$

(taking into account that $\sum_{h=1}^{n}((X_h^P)_{i,j} - \hat{\mu}_{i,j}) = 0$; recall that $\hat{\mu}_{i,j} = \frac{1}{n}\sum_{h=1}^{n}(X_h^P)_{i,j}$). Owing to these results, we only need to minimize the second addend:

$$\begin{aligned}
\sum_{h=1}^{n} d^2_{OPA}(X_{n+1}, X_h^P) &= \sum_{h=1}^{n}\sum_{i=1}^{k}\sum_{j=1}^{m}(\hat{\mu}_{i,j} - \beta(X_{n+1})_{i,*} \cdot \Gamma_{*,j} - 1_k\gamma^T)^2 \\
&= n \cdot \sum_{i=1}^{k}\sum_{j=1}^{m}(\hat{\mu}_{i,j} - \beta(X_{n+1})_{i,*} \cdot \Gamma_{*,j} - 1_k\gamma^T)^2 \\
&= n \cdot \|\hat{\mu} - \beta X_{n+1}\Gamma - 1_k\gamma^T\|^2 \\
&= n \cdot d^2_{OPA}(X_{n+1}, \hat{\mu}).
\end{aligned}$$

As a result, minimizing $\sum_{h=1}^{n} d^2_{OPA}(X_{n+1}, X_h^P)$ is equivalent to minimizing $d^2_{OPA}(X_{n+1}, \hat{\mu})$. □

3.2 Registration to the median shape

Although the registration to the Procrustes sample mean shape is a good option in order to align an out-of-sample observation with the set of Procrustes coordinates, it might happen that the Procrustes sample is very heterogeneous, and that, as a result, the registration obtained does not 'resist' the possible existence of outliers. Therefore, it is necessary to introduce a more robust registration method which is the registration to the Procrustes sample median shape.

This method is based on the definition of the median curve in the field of functional data (which is an adaptation of the definition of the median of a random variable), where a curve is considered to be the (sample) median when it maximizes the depth within a given set of curves defined on the same domain (López-Pintado, Romo, 2006). Intuitively, the median curve is the curve most surrounded in a sample of curves, i.e., it is the most central curve. In our particular case, given a sample of configuration matrices $X_1, ..., X_n$, the median configuration matrix is considered to be a matrix belonging to the sample, i.e., X_{h_0} for $h_0 \in \{1, ..., n\}$, such that its landmarks (as a whole) are the most surrounded ones within all the possible configurations of landmarks belonging to the sample. On the other hand, it has to be taken into account that this sample median is not the arithmetic median of each of the sample landmarks (the arithmetic median is not a shape belonging to the sample).

Let us take a look at the following example in order to understand this concept:

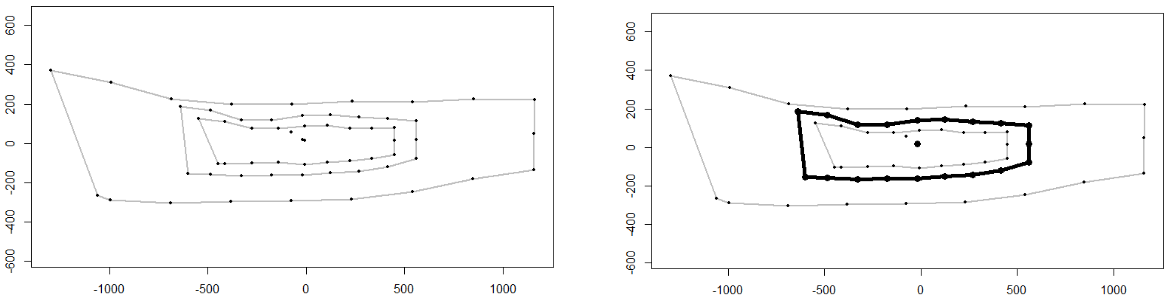


Figure 3.2: Raw landmarks' coordinates of 3 arms. The sample median (black arm) is represented on the right figure. Notice how, while the median arm is surrounded by the other 2 arms, these 2 arms are each surrounded by the median and, hence, they are less deep.

Having understood what the median shape represents, we are able to introduce the following definition:

Definition 3.2.1. *Let $X_1^P, ..., X_n^P$ the Procrustes coordinates of $X_1, ..., X_n \in M_{k\times m}(\mathbb{R})$ and let $X_{n+1} \in M_{k\times m}(\mathbb{R})$ be an out of sample observation. Let also $X_{h_0}^P$ be the deepest observation among the Procrustes coordinates, i.e., the sample median Procrustes configuration matrix. Then, the* ***registration of X_{n+1} to the (Procrustes sample) median shape*** *is given by*

$$X_{n+1}^R = \hat{\beta} X_{n+1} \hat{\Gamma} + 1_k \hat{\gamma}^T,$$

where $\hat{\beta}, \hat{\Gamma}, \hat{\gamma}$ are the minimizing similarity parameters estimated in equation (2.1) after registering X_{n+1} to $X_{h_0}^P$ with full OPA.

The sample median represented in Figure 3.2 is not the Procrustes sample median. Hence, we need to introduce a different example in order to understand how does the registration to the Procrustes sample median really works:

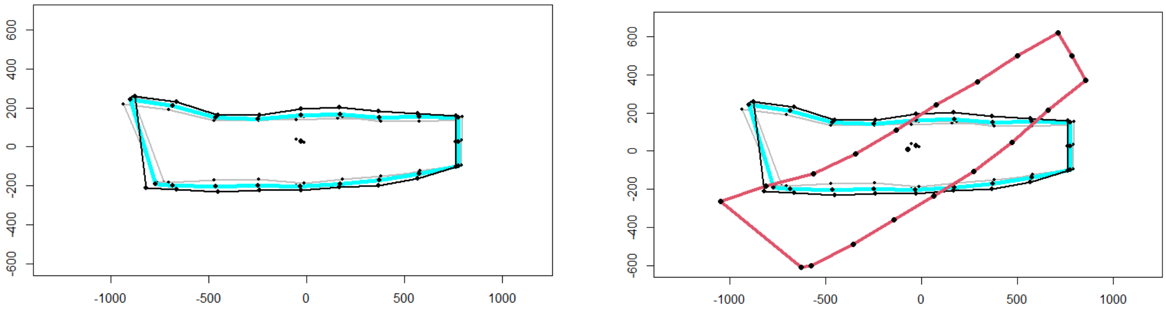


Figure 3.3: Procrustes coordinates of the arms given in Figure 3.2 along with the Procrustes sample mean (blue) and Procrustes sample median (black). On the right hand Figure, an out-of-sample observation (red) is represented.

Taking into account that the sample mean and the sample median are not obtained in the same way, then the registrations of an out-of-sample observation to both shapes will be different. In this particular example, the registrations obtained are very close (but still they are not the same):

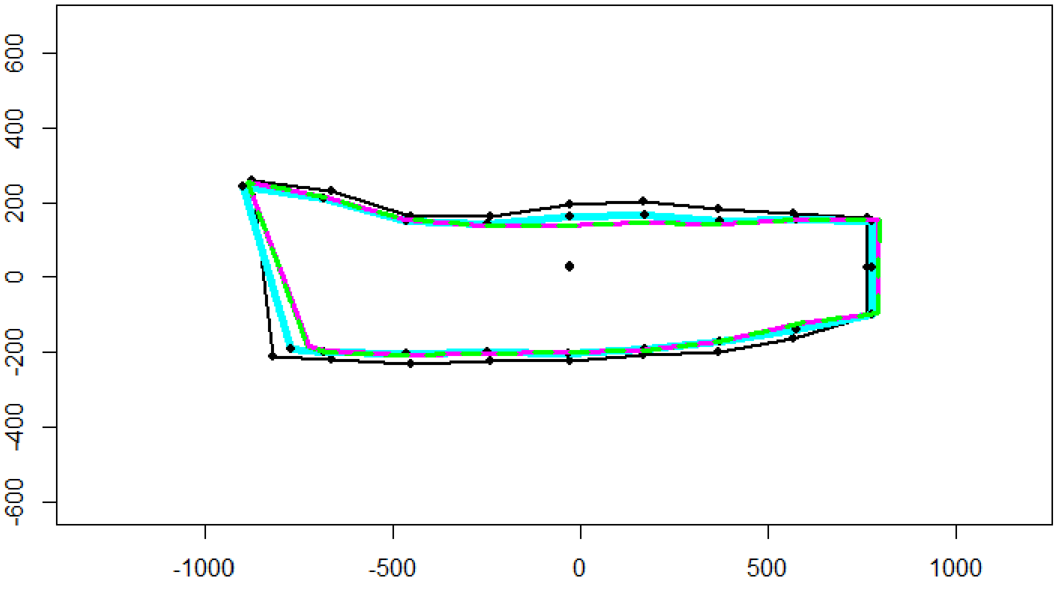


Figure 3.4: Registration (pink) of the out-of-sample arm given in Figure 3.3 to the Procrustes sample mean (blue), and registration (green) of the same out-of-sample arm to the Procrustes sample median (black).

Chapter 4

Case study

In this chapter we will put into practice the techniques introduced throughout this Thesis with the aim of creating a classifier based on a real sample of children. This said, we have a dataset of Senegalese children having 957 observations, where each child has the following information associated:

- Left arm raw bi-dimensional landmark coordinates: in total, there are 40 variables summarizing the information of 20 landmarks. These coordinates have been obtained by SAM photo app after translation of the respective left arrm photograph has been performed. 16 of these landmarks are semi-landmarks obtained a posteriori. Additionally, some of these landmarks have a scientific meaning as follows: landmark 1 is located on the shoulder, landmark 4 is located on the armpit, landmarks 17, 2 and 18 are located on the elbow and landmarks 19, 3 and 20 are located on the wrist. To visualize the resulting configuration of landmarks, we can take a look at any of the Figures given throughout Chapter 2 (notice that all of the Figures in this Thesis have been obtained from this sample dataset).
- agemons : age (months).
- sex : whether male, 1, or female, 2.
- weight (kg).
- height (cm).
- muac (cm).
- measure : whether the height of the child has been measured while the child was lying down, 'l', or standing up, 'h'.
- zlen : z-score of the length for age (obtained from a reference sample).
- zwei : z-score of the weight for age (obtained from a reference sample).
- zwfl : z-score of the weight for height (obtained from a reference sample).
- zbmi : z-score of the body max index for age (obtained from a reference sample).
- zac : z-score of MUAC for age (obtained from a reference sample).
- class_muac : classification of a child (whether NOR, RIS, MAM or SAM) using MUAC criteria (previously introduced in the Introduction).
- class_wfl : classification of a child (whether NOR, RIS, MAM or SAM) using WFL criteria (previously introduced in the Introduction).

- class_global : classification of a child (whether NOR, RIS, MAM or SAM) taking the most restrictive classification between class_muac and class_wfl (e.g., for a given child belonging to NOR and MAM at the same time, then the nutritional status is set to the worst case, MAM; this way, we avoid missing a child who might be undernourished and that one of the criteria did not detect it).

Additionally, there exist 2 partition datasets within the 957 observations. The first partition is composed by the first 570 observations while the other partition is composed by the remaining 387 observations. The main difference between both partition is basically the left arm position held by children during the photo shoot. More precisely, children from the first partition had their arm captured in a straighter position than those from the second partition (this can be checked in the Figures shown in the next section).

This said, the population of interest is given by the children from Senegal between 5 and 59 months.

On the other hand, taking into account that environmental factors (like, for instance, the altitude; Pawson, Huicho, Muro, Pacheco, 2019) and genetic factors might have an effect on the biological adaption of human populations and, hence, in human proportions, then it is very likely that any classifier that we fit in the dataset of Senegalese children is biased to this population (i.e., it is biased to the particular arm shape that Senegalese children have). As a result, this might result in a worse performance of the classifiers when extrapolating to a different population (whose arm shapes might differ).

4.1 Code

All the code developed in this Thesis is available on the following repository from GitHub called shapesSL :

`https://github.com/alvaroperezromero/shapesSL.git`

In total, 5 functions have been developed:

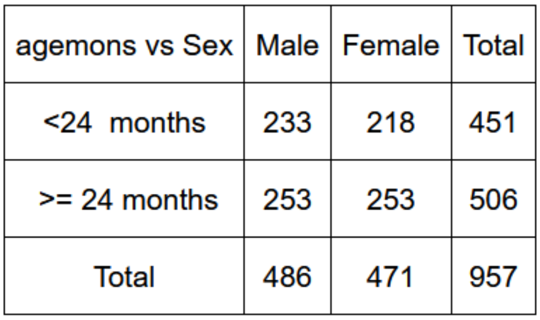
- `shapesOutliers`: it finds outliers in a given set of shapes.
- `shapesRegistration` it performs registration to the Procrustes sample median or Procrustes sample mean of a given set of out-of-sample shapes.
- `shapesClassification` for a training-test partition, it fits a classification model in order to check the performance in the test set. This function uses the previous 2 ones.
- `shapesClassification_2` analogous to shapesClassification but it checks the average performance in a set of partitions (obtained either by Cross Validation or simply random splits).
- `SAM_final` this function uses all the previous 4 functions in order to check the performance in our study case sample dataset.

Furthermore, the scripts developed use the following packages: shapes , pracma , plyr , roahd , geomorph , caret , foreach , doParallel .

4.2 Descriptive analysis

We are only going to focus on the main variables

To begin with, we can see that the distribution of agemons and sex is balanced. Let us take a look at the next Figure (taking into account that the maximum of agemons is 60.4 months):



agemons vs Sex	Male	Female	Total
<24 months	233	218	451
>= 24 months	253	253	506
Total	486	471	957

Figure 4.1: Age in months (2 levels) vs sex of Children.

Now, regarding to measure, 507 children have had their length measured while they were standing up and 450 of them while they were lying down.

With respect to weight and height, if a child is undernourished or is at risk of being undernourished, then the z-score of the weight for age, zwfl, decreases, while if the child is well nourished the opposite happens. Regarding to muac and zwfl distributions vs class_global, it is clear that WFL criteria is more restrictive than MUAC criteria taking into account that zwfl distribution overlaps less among the different levels. Also, given that the Body Max Index of a child is obtained in a similar way than the weight for length, then zwfl and zbmi will be correlated and, hence, the similar densities are expected. Let us check this in the following Figure:

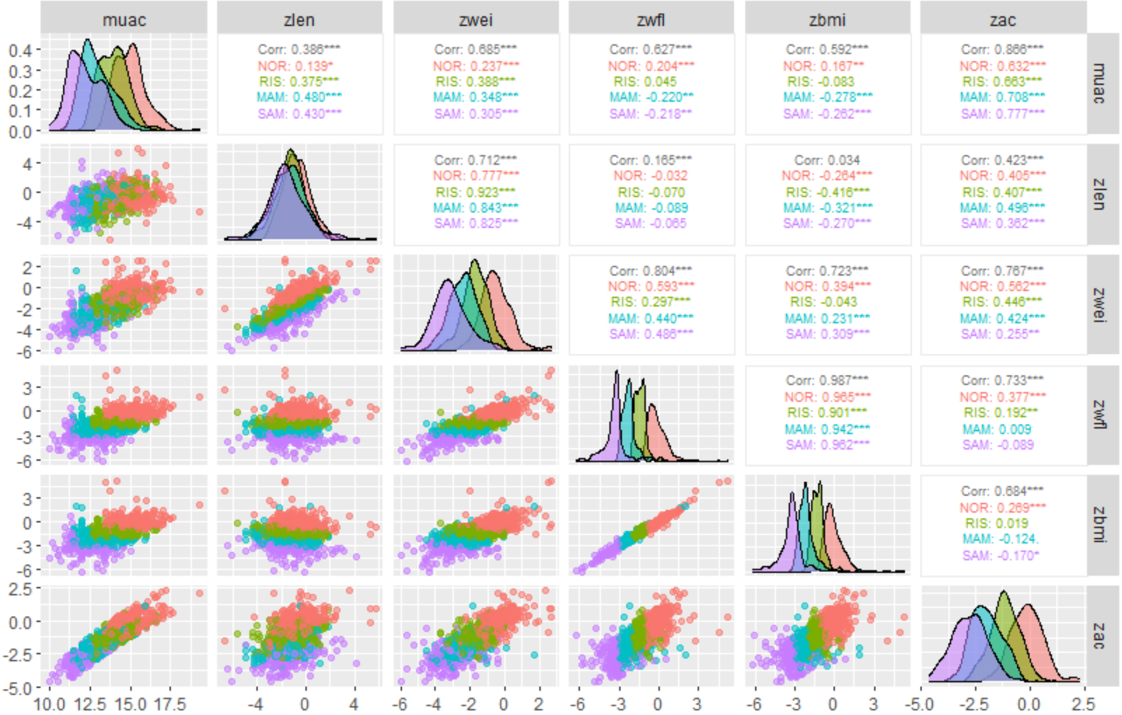


Figure 4.2: densities of muac, zlen, zwei, zwfl, zbmi, zac vs class_global

On the other hand, 8.1% of the times class_global value is obtained because MUAC criteria is more restrictive than WFL criteria. Also, 46% of the times class_global is owing to MUAC criteria is less restrictive than WFL criteria. And, finally, 45.9% of the times both criteria match in their classifications. This said, it seems that WFL criteria is a more restrictive criteria. Additionally, we can take a look at the distribution of children in the different nutritional status within the possible partitions:

	NOR	RIS	MAM	SAM	NOR + RIS	MAM+SAM	Total
1st partition	138	142	147	143	280	290	570
2nd partition	167	138	62	20	305	82	387
All children	305	280	209	163	585	372	970

Figure 4.3: Nutritional status distribution by partition.

4.3 Procrustes analysis

Taking into account that the dataset is partitioned, we might want to analyze the shapes separately for each of the partitions:

- **First partition**

  If we obtain the Procrustes coordinates of all the 570 observations belonging to the first partition, then the following is observed depending on the nutritional levels considered:

  - (NOR, RIS, MAM, SAM): it seems that, although the bi-dimensional points overlap within all the 20 possible landmarks, differences are observed depending on the nutritional status. More precisely, for any given landmark (except for those located on the wrist, i.e., landmarks 19, 3 and 20), the location of the 570 bi-dimensional Procrustes points associated to that landmark is different depending on the level, in the sense that, those bi-dimensional points belonging to SAM group tend to narrow the shape of the arm, and those bi-dimensional points belonging to NOR group tend to widen the shape of the arm. Intuitively, this makes sense given that undernourished children are expected to have an arm shape which tends to narrow in the anatomical spots represented by the landmarks, while it is expected the opposite for those who are better nourished. In order to check this, let us take a look at the following Figure:

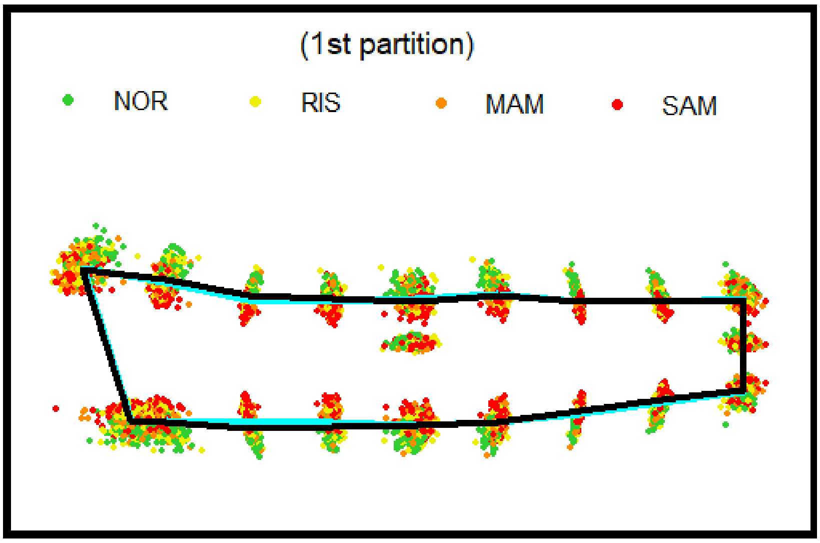


Figure 4.4: Procrustes coordinates (first partition - 4 groups) along with Procrustes sample mean (blue) and Procrustes sample median (black).

- (NOR - RIS, MAM - SAM): If 2 nutritional status are rather considered, then differences among them are even clearer now. Besides, analogous conclusions can be made regarding to the relationship between the location of the landmarks and the nutritional status. Nonetheless, it seems that now we can appreciate differences regarding to the shape of the wrist (while in the study of the 4 classes they were not observed). This said, it seems that, owing to this better separation between classes, models will perform better than when 4 classes are considered. Let us take a look at the following Figure:

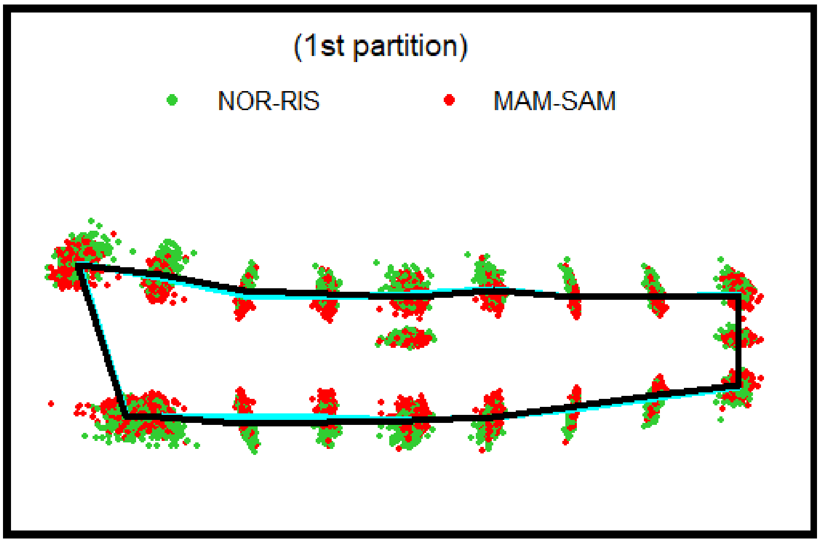


Figure 4.5: Procrustes coordinates (first partition - 2 groups) along with Procrustes sample mean (blue) and Procrustes sample median (black).

- **Second partition**
  If we obtain the Procrustes coordinates of all the 387 observations belonging to the second partition, then different plots are obtained. In particular, most of the arms plotted belong to NOR and RIS (take into account that class_global is unbalanced; see Figure 4.3). Furthermore, we need to bear in mind that the arm position captured differs from that of the first partition (in order to check this, we just need to take a look at the mean shapes obtained in partitions 1 and 2) :

  - (NOR, RIS, MAM, SAM): Although there exist just 20 children out of 387 children belonging to *SAM* group, it can still be checked how there exist some of them which do not overlap. Hence, this might differences in shape. Also, in this case, the differences between NOR and RIS levels are appreciated better. Let us take a look at the following Figure:

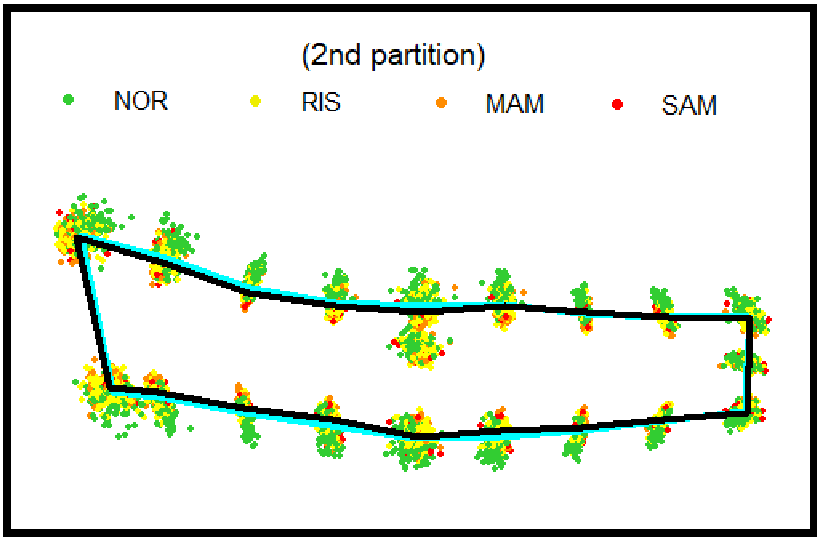


Figure 4.6: Procrustes coordinates (second partition - 4 groups) along with Procrustes sample mean (blue) and Procrustes sample median (black).

- (NOR - RIS, MAM - SAM): If 2 nutritional status are considered, then children belonging to NOR-RIS class will mainly be represented:

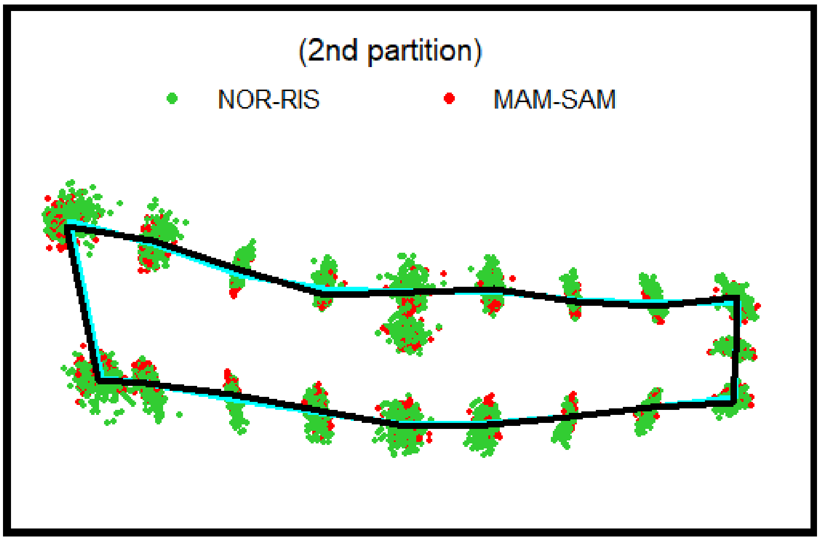


Figure 4.7: Procrustes coordinates (second partition - 2 groups) along with Procrustes sample mean (blue) and Procrustes sample median (black).

- **All children**
  Although now we have joined the first and second partition (which might introduce noise taking into account that the mean shape in both cases are different), analogous conclusions can be made as in the first partition.
  - (NOR, RIS, MAM, SAM): If we take a look at the following Figure, there also seem to be differences between nutritional status.

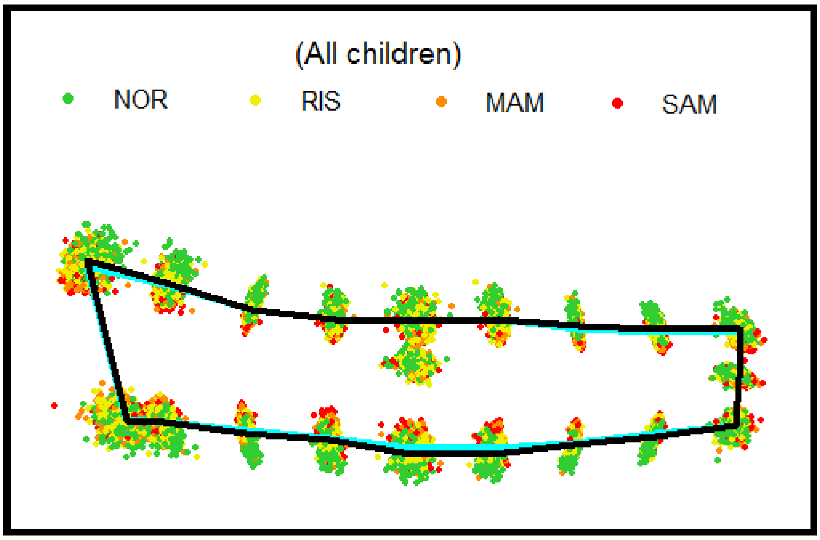


Figure 4.8: Procrustes coordinates (all children - 4 groups) along with Procrustes sample mean (blue) and Procrustes sample median (black).

- (NOR - RIS, MAM - SAM): The next Figure might be misleading in the sense that we might think that a classification model will perform worse (because of the overlapping) than in the analogus example of the first partition. Still, it needs to be taken into account that, given that partitions 1 and 2 have been joined, classes NOR-RIS is over-represented:

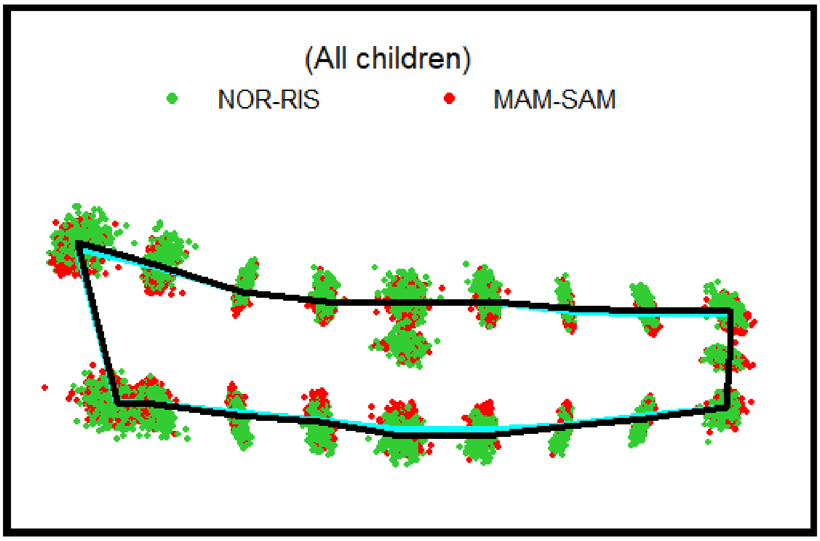


Figure 4.9: Procrustes coordinates (all children - 2 groups) along with Procrustes sample mean (blue) and Procrustes sample median (black).

4.3.1 Outliers

Before we even started making a classifier in our dataset, we used shapesOutliers in order to check whether there existed outliers shapes originated by SAM photo app translation's method. As a result, the following outliers were obtained and fixed:

- Observations 574 and 794 needed to be rotated 180 degrees:

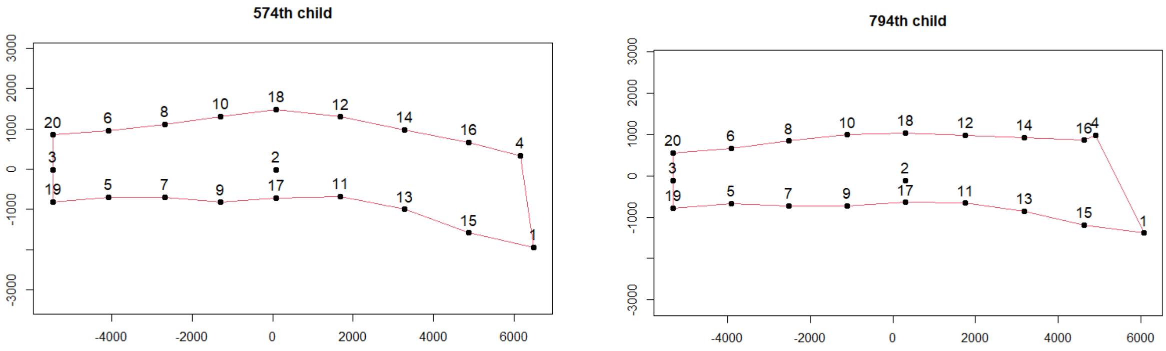


Figure 4.10: Raw coordinates obtained by SAM photo app.

- Landmarks 4 and 16 (4th and 16th rows of the configuration matrices) needed to be exchanged in children $115, 119, 123, 128, 132, 175$:

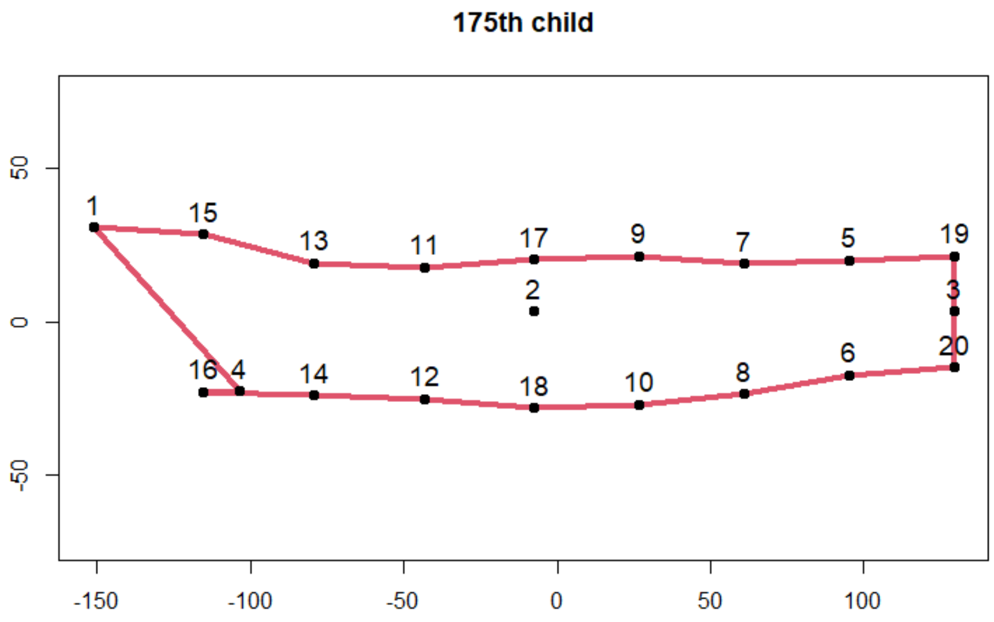


Figure 4.11: Raw coordinates obtained by SAM photo app.

- Analogously, landmarks $11, 15$ and landmarks $16, 14, 12$ needed to be fixed for children 933 to 957:

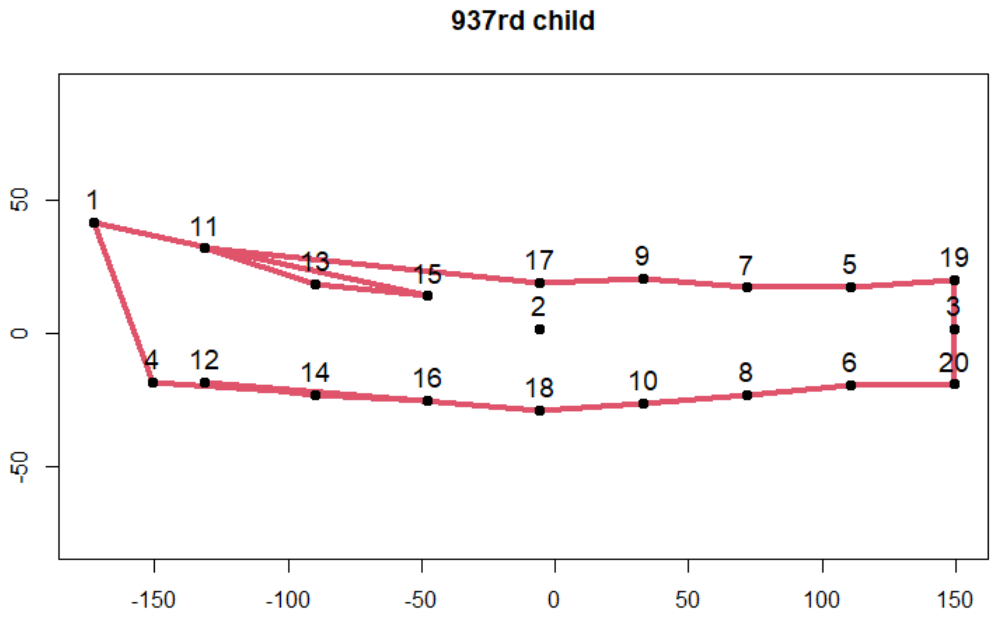


Figure 4.12: Raw coordinates obtained by SAM photo app.

- Child 268 was removed from the dataset:

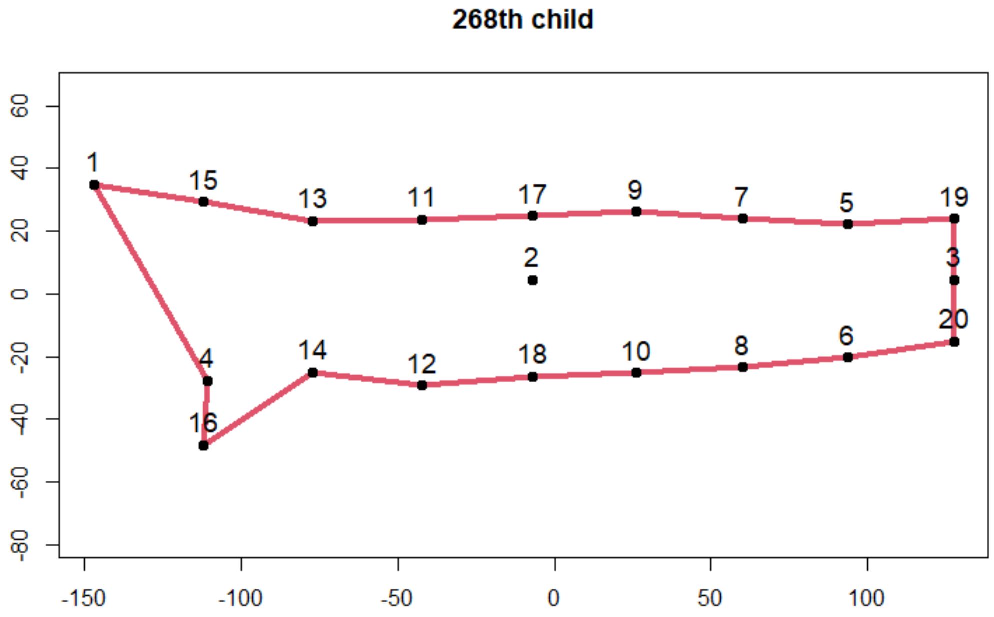


Figure 4.13: Raw coordinates obtained by SAM photo app.

4.3.2 Classification

The classification results can be checked in the Appendix.

Once the data has been pre-processed and the outliers mentioned in the previous section have been fixed, now it is time to check how SAM photo diagnosis app might perform in reality. The outline followed in order to estimate this is as follows (the function used is `SAMfinal`):

1. Select either the first partition or all the dataset.
2. Select either children below 24 months or children above or equal than 24 months.
3. Select either 4 classes or 2 classes in the nutritional status.
4. For the resulting data from the previous 3 steps, obtain a 20-fold set of training-test partitions (although ideally leave-one-out CV is the method that better simulates the classification of an out-of-sample observation, 20-CV performance will be very similar). Then, for each partition do as follows (with parallelization):
5. Remove outliers from the training set considering the Mahalanobis distances of each of the Procrustes training observations to the training Procrustes sample mean. Thus, observations which distance is greater than $q3 + 1.5IQR$ (being IQR the interquantile range) are removed.
6. Obtain the Procrustes coordinates of the new set of training observations without the outliers and save them. Then, perform Allometry and obtain the final set of training coordinates.
7. For each registration method (whether mean or median) register the shapes belonging to the test set to the Procrustes set saved using the registration method. Then, obtain the final coordinates using the Allometry parameters obtained before.

8. Fit a model (between KNN, LDA or Decision Tree) and obtain the predicted classes.
9. After 20-folds CV is finished, we return the Confusion Matrix of the actual classes vs the predicted classes and check the sensitivities, specificites and FPR.

Chapter 5

Conclusions

- There exists too much overlapping when 4 classes are considered. Therefore, the accuracies are not very good. Owing to this, it was decided to decrease the number of classes to just 2: NOR - RIS, MAM - SAM.
- Given that the classifiers trained might be biased to the population of Senegalese children between 5 and 59 months, it would be a good idea to check the performance on a different population of children (maybe from other country).
- Although individually the models do not perform very well (specially when classifying in the 4-class case), they can be combined in order to make the final decision regarding to the nutritional status of a child.
- In further work, we can try to remove outliers by triangulations between landmarks (for instance, the triangulation of 1, 2 and 4 landmarks. There are 4 possible triangulations. On the other hand, taking into account that SAM photo diagnosis app might not properly translate an out-of-sample photograph into a set of landmarks, then it would be a good idea to implement in the app an outlier detection method for the out-of-sample observations.
- The reason why the mean and median registration methods perform relatively close in our case has to do with the fact that all the training shapes have been standardized before (the Procrustes coordinates) and also that all the shapes represent the same kind of shape, i.e., an arm. However, if there existed observations that represented different kind of shapes, and the sample was heterogeneous, then maybe the median shape registration method would be better.
- Actually, the goal is not to obtain the best model having the best accuracy, but to train models which are computationally efficient and that can be implemented in SAM photo diagnosis app. We need to bear in mind that this app is designed to work offline, so it is better to consume low memory space. In this regard, the models selected are simple classification rules.
- If the size was captured 100% exact, an study of size-and-shape could be performed, i.e., a study where we consider that 2 forms are equal when they share the same size and shape.

Bibliography

[1] Action Against Hunger. (n.d.). Severe Acute Malnutrition (SAM) Photo Diagnosis App® Project, accessed 3 Sep 2022. Retrieved from `https://knowledgeagainsthunger.org/research/prevention/severe-acute-malnutrition-sam-photo-diagnosis-app-project/`.

[2] Adams, C., James Rohlf, F., Slice, Dennis E. (2004). Geometric morphometrics: Ten years of progress following the 'revolution', *Italian Journal of Zoology*, 71:1, 5-16. `https://doi.org/10.1080/11250000409356545`.

[3] Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions. *Statistical Science*, 1, 181–242.

[4] Bookstein, F. L. (1991). Morphometric tools for landmark data: Geometry and biology. *Cambridge, United Kingdom: Cambridge University Press,* 256–35.

[5] Bookstein, F. L. (1997). Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape. *Medical Image Analysis*, 1(3), 225–243.

[6] de Onis, M., Habicht, J. D. (1996). Anthropometric reference data for international use: Recommendations from a world health organization expert committee. *American Journal of Clinical Nutrition*, 64, 650–658

[7] Dryden, I. L. , Mardia, K.V. (2016). Statistical Shape Analysis: with Applications in R (2nd. ed.). *Wiley.*

[8] Gower, J. C. (1975). Generalized Procrustes analysis. *Psychometrika*, 40: 33–50. page 125, 133, 136, 138.

[9] Grellety, E., Golden, M. H. (2016). Weight-for-height and mid-upper-arm circumference should be used independently to diagnose acute malnu- trition: Policy implications. *BMC Nutrition*, 2, 10.

[10] Helmert, F. R. (1876), Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Beobachtungsfehlers directer Beobachtungen gleicher Genauigkeit, *Astronom. Nachr.*, 88 115–132.

[11] Hurley, J. R., Cattell, R. B.. (1962). The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Computers in Behavioral Science.*

[12] Kendall, D. G. (1977). The diffusion of shape. *Advances in Applied Probability*, 9: 428–430. page 1, 33, 351.

[13] Kendall, D. G. (1984). Shape manifolds, Procrustean metrics and complex projective spaces. *Bulletin of the London Mathematical Society*, 16: 81–121. page xxi.

[14] López-Pintado, S., Romo, J. (2006). On the concept of depth for functional data, *DES - Working Papers. Statistics and Econometrics.* WS ws063012, Universidad Carlos III de Madrid. Departamento de Estadística.

[15] López-Pintado, S., Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718 – 734.

[16] López-Pintado, S., Romo, J. (2011). A half-region depth for functional data. *Computational Statistics Data Analysis*, 55(4):1679–1695. [p6, 7].

[17] Medialdea L., Bazaco C., D'Angelo del Campo MD., et al. (2019). Describing the children's body shape by means of Geometric Morphometric techniques. *Am J Phys Anthropol.* 168:651–664. `https://doi.org/10.1002/ajpa.23779`

[18] Medialdea, L., Bogin, B., Thiam, M. et al. (2021). Severe acute malnutrition morphological patterns in children under five. *Sci Rep.* 11, 4237. `https://doi.org/10.1038/s41598-021-82727-x`

[19] Myatt, M., Duffield, A., Seal, A., Pasteur, F. (2009). The effect of body shape on weight-for-height and mid-upper arm circumference based case definitions of acute malnutrition in Ethiopian children. *Annals of Human Biology*, 36(1), 5–20.

[20] Rohlf, F. J., Slice, D. (1990). Extensions of the Procrustes method for the optimal superimposition of landmarks. Systematic Zoology, 39: 40–59. page 277, 336, 338.

[21] Sandoval Oyanedel, C., Diaz, A., Manríquez, G. (2021). Assessing cervical spine and craniofacial morphology in Class II and Class III malocclusions: A geometric morphometric approach. *Cranio : the journal of craniomandibular practice.* 1-11.

[22] WikiHow (2022). Pull the tape snug against the child's arm, accessed 4 Sep 2022. Retrieved from `https://www.wikihow.com/Measure-Mid-Arm-Circumference#/Image:Measure-Mid-Arm-Circumference-Step-4.jpg`.

[23] World Health Organization (WHO). (2013). Pocket Book of Hospital Care for Children: Guidelines for the Management of Common Childhood Illnesses (2nd. ed.). *World Health Organization.*

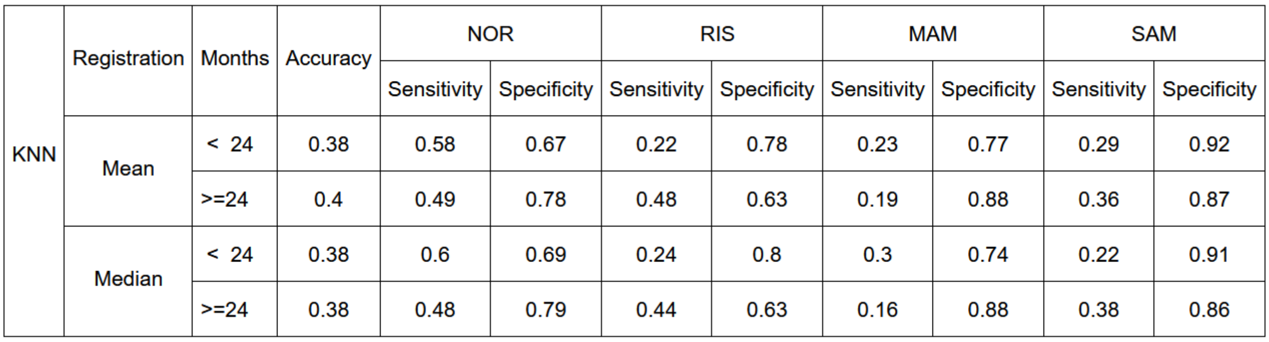
[24] World Health Organization (WHO). (2020). Malnutrition, accessed 3 Sep 2022. Retrieved from `https://www.who.int/news-room/fact-sheets/detail/malnutrition`.

Appendix A

Classification results

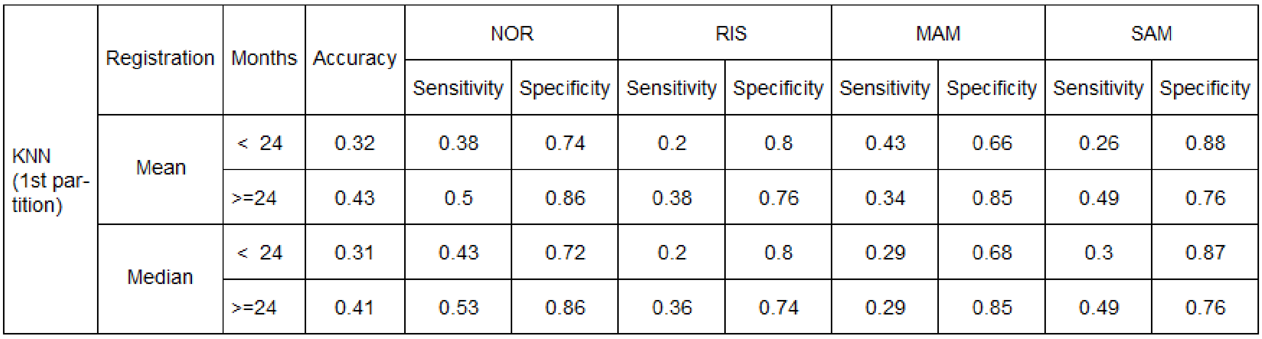
A.1 4 classes: NOR, RIS, MAM, SAM

A.1.1 KNN



	Registration	Months	Accuracy	NOR		RIS		MAM		SAM	
				Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
KNN	Mean	< 24	0.38	0.58	0.67	0.22	0.78	0.23	0.77	0.29	0.92
		>=24	0.4	0.49	0.78	0.48	0.63	0.19	0.88	0.36	0.87
	Median	< 24	0.38	0.6	0.69	0.24	0.8	0.3	0.74	0.22	0.91
		>=24	0.38	0.48	0.79	0.44	0.63	0.16	0.88	0.38	0.86

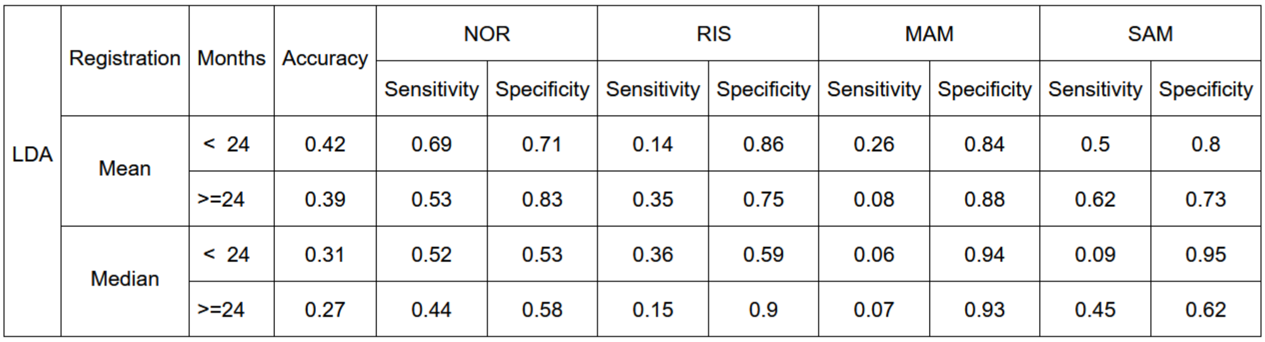
Figure A.1: KNN results (all children).



	Registration	Months	Accuracy	NOR		RIS		MAM		SAM	
				Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
KNN (1st partition)	Mean	< 24	0.32	0.38	0.74	0.2	0.8	0.43	0.66	0.26	0.88
		>=24	0.43	0.5	0.86	0.38	0.76	0.34	0.85	0.49	0.76
	Median	< 24	0.31	0.43	0.72	0.2	0.8	0.29	0.68	0.3	0.87
		>=24	0.41	0.53	0.86	0.36	0.74	0.29	0.85	0.49	0.76

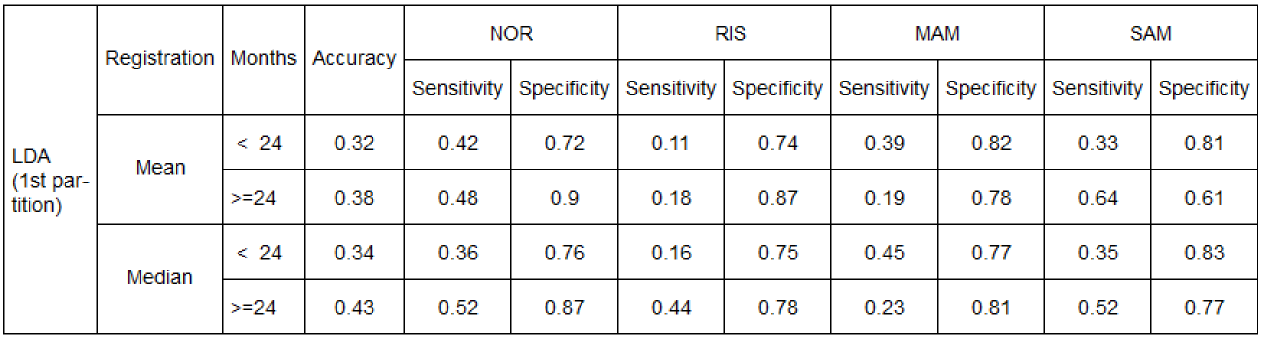
Figure A.2: KNN results (first partition of children).

A.1.2 Linear Discriminant Analysis



	Registration	Months	Accuracy	NOR		RIS		MAM		SAM	
				Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
LDA	Mean	< 24	0.42	0.69	0.71	0.14	0.86	0.26	0.84	0.5	0.8
		>=24	0.39	0.53	0.83	0.35	0.75	0.08	0.88	0.62	0.73
	Median	< 24	0.31	0.52	0.53	0.36	0.59	0.06	0.94	0.09	0.95
		>=24	0.27	0.44	0.58	0.15	0.9	0.07	0.93	0.45	0.62

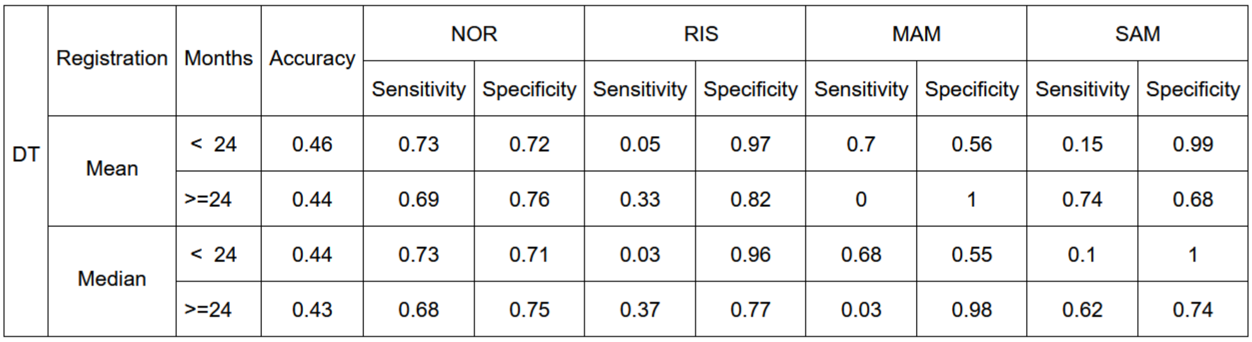
Figure A.3: Linear Discriminant Analysis results (all children).



	Registration	Months	Accuracy	NOR		RIS		MAM		SAM	
				Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
LDA (1st partition)	Mean	< 24	0.32	0.42	0.72	0.11	0.74	0.39	0.82	0.33	0.81
		>=24	0.38	0.48	0.9	0.18	0.87	0.19	0.78	0.64	0.61
	Median	< 24	0.34	0.36	0.76	0.16	0.75	0.45	0.77	0.35	0.83
		>=24	0.43	0.52	0.87	0.44	0.78	0.23	0.81	0.52	0.77

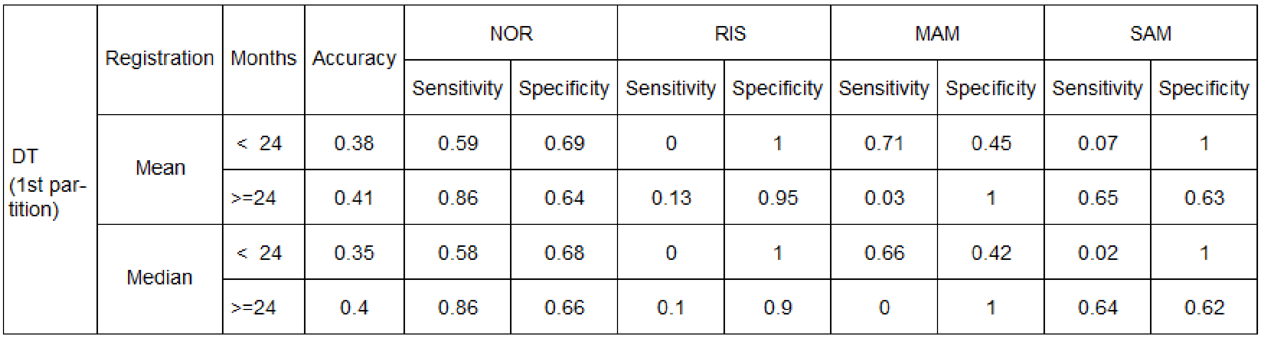
Figure A.4: Linear Discriminant Analysis results (first partition of children).

A.1.3 Decision Tree



	Registration	Months	Accuracy	NOR		RIS		MAM		SAM	
				Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
DT	Mean	< 24	0.46	0.73	0.72	0.05	0.97	0.7	0.56	0.15	0.99
		>=24	0.44	0.69	0.76	0.33	0.82	0	1	0.74	0.68
	Median	< 24	0.44	0.73	0.71	0.03	0.96	0.68	0.55	0.1	1
		>=24	0.43	0.68	0.75	0.37	0.77	0.03	0.98	0.62	0.74

Figure A.5: Decision tree results (all children).

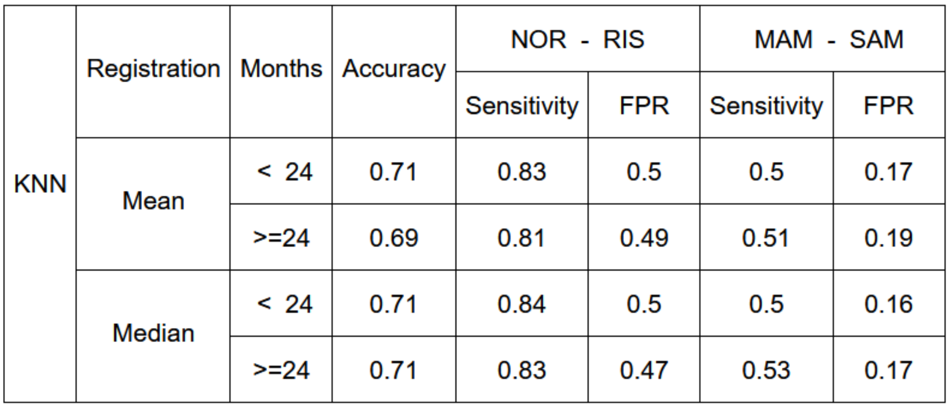


	Registration	Months	Accuracy	NOR		RIS		MAM		SAM	
				Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
DT (1st partition)	Mean	< 24	0.38	0.59	0.69	0	1	0.71	0.45	0.07	1
		>=24	0.41	0.86	0.64	0.13	0.95	0.03	1	0.65	0.63
	Median	< 24	0.35	0.58	0.68	0	1	0.66	0.42	0.02	1
		>=24	0.4	0.86	0.66	0.1	0.9	0	1	0.64	0.62

Figure A.6: Decision tree results (first partition of children).

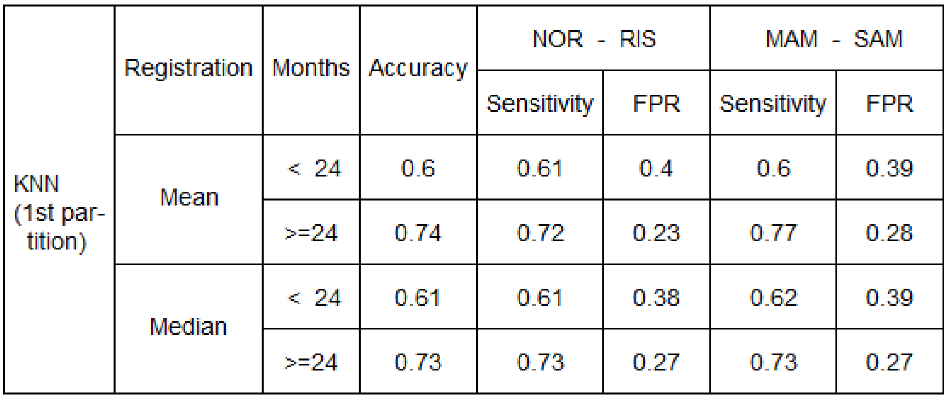
A.2 2 classes: NOR - RIS, MAM - SAM

A.2.1 KNN



	Registration	Months	Accuracy	NOR - RIS		MAM - SAM	
				Sensitivity	FPR	Sensitivity	FPR
KNN	Mean	< 24	0.71	0.83	0.5	0.5	0.17
		>=24	0.69	0.81	0.49	0.51	0.19
	Median	< 24	0.71	0.84	0.5	0.5	0.16
		>=24	0.71	0.83	0.47	0.53	0.17

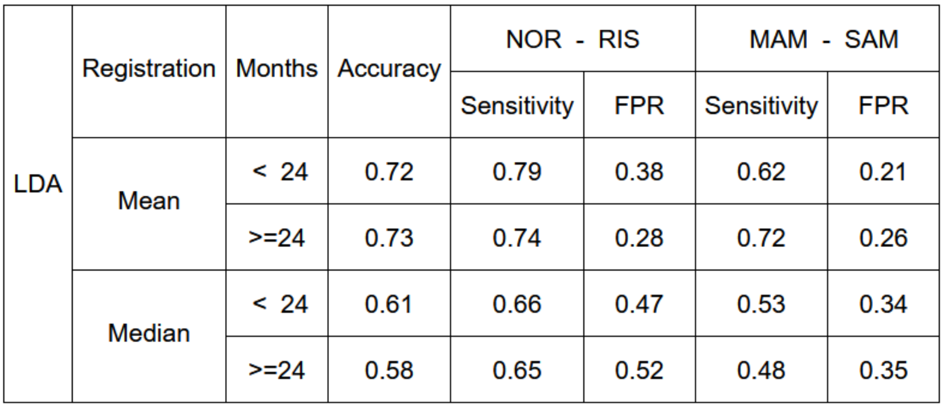
Figure A.7: KNN results (all children). FPR stands for False Positive Rate.



	Registration	Months	Accuracy	NOR - RIS		MAM - SAM	
				Sensitivity	FPR	Sensitivity	FPR
KNN (1st partition)	Mean	< 24	0.6	0.61	0.4	0.6	0.39
		>=24	0.74	0.72	0.23	0.77	0.28
	Median	< 24	0.61	0.61	0.38	0.62	0.39
		>=24	0.73	0.73	0.27	0.73	0.27

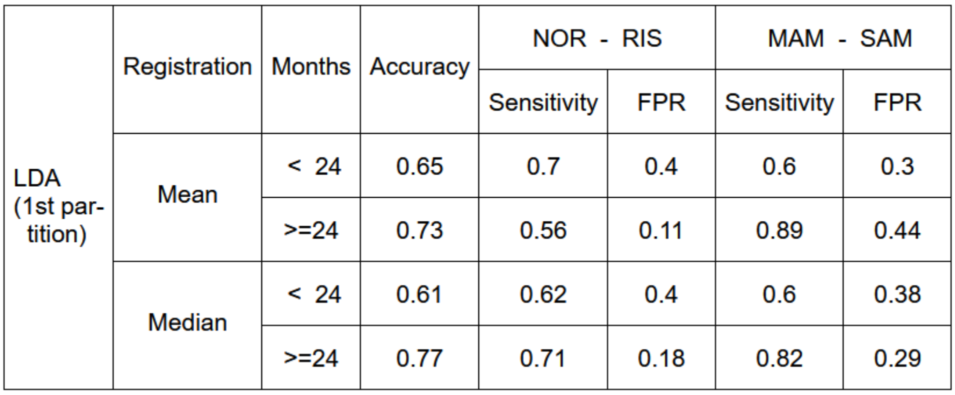
Figure A.8: KNN results (first partition of children). FPR stands for False Positive Rate.

A.2.2 Linear Discriminant Analysis



	Registration	Months	Accuracy	NOR - RIS		MAM - SAM	
				Sensitivity	FPR	Sensitivity	FPR
LDA	Mean	< 24	0.72	0.79	0.38	0.62	0.21
		>=24	0.73	0.74	0.28	0.72	0.26
	Median	< 24	0.61	0.66	0.47	0.53	0.34
		>=24	0.58	0.65	0.52	0.48	0.35

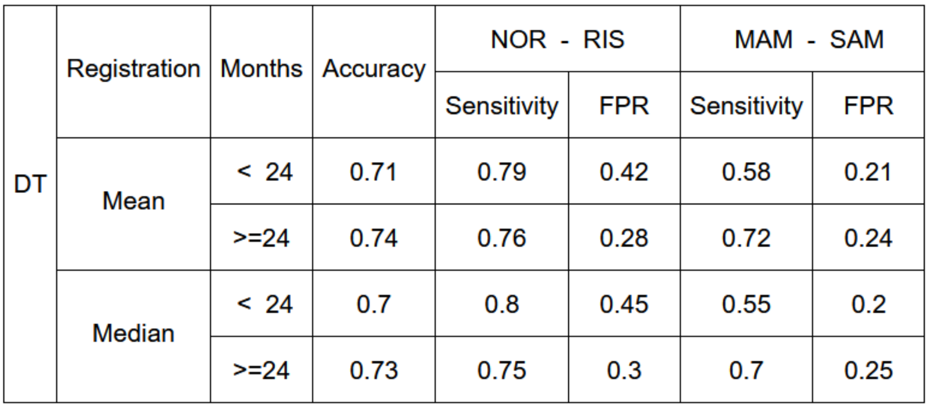
Figure A.9: Linear Discriminant Analysis results (all children). FPR stands for False Positive Rate.



	Registration	Months	Accuracy	NOR - RIS		MAM - SAM	
				Sensitivity	FPR	Sensitivity	FPR
LDA (1st partition)	Mean	< 24	0.65	0.7	0.4	0.6	0.3
		>=24	0.73	0.56	0.11	0.89	0.44
	Median	< 24	0.61	0.62	0.4	0.6	0.38
		>=24	0.77	0.71	0.18	0.82	0.29

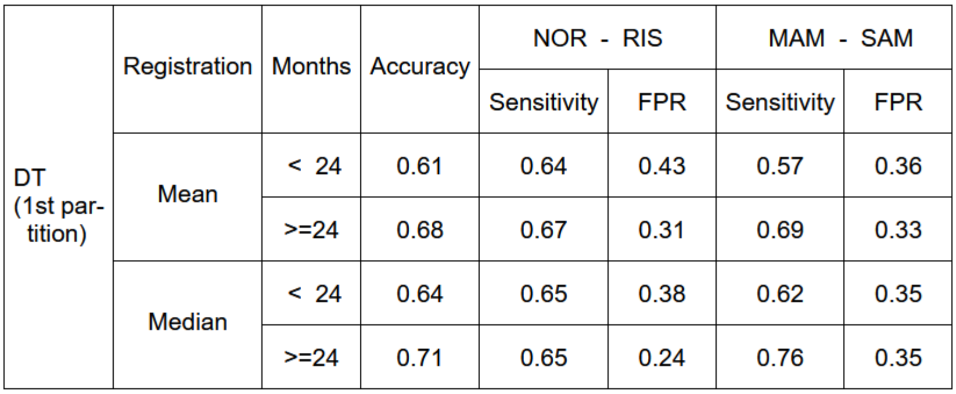
Figure A.10: Linear Discriminant Analysis results (first partition of children). FPR stands for False Positive Rate.

A.2.3 Decision Tree



	Registration	Months	Accuracy	NOR - RIS		MAM - SAM	
				Sensitivity	FPR	Sensitivity	FPR
DT	Mean	< 24	0.71	0.79	0.42	0.58	0.21
		>=24	0.74	0.76	0.28	0.72	0.24
	Median	< 24	0.7	0.8	0.45	0.55	0.2
		>=24	0.73	0.75	0.3	0.7	0.25

Figure A.11: Decision tree results (all children). FPR stands for False Positive Rate.



	Registration	Months	Accuracy	NOR - RIS		MAM - SAM	
				Sensitivity	FPR	Sensitivity	FPR
DT (1st partition)	Mean	< 24	0.61	0.64	0.43	0.57	0.36
		>=24	0.68	0.67	0.31	0.69	0.33
	Median	< 24	0.64	0.65	0.38	0.62	0.35
		>=24	0.71	0.65	0.24	0.76	0.35

Figure A.12: Decision tree results (first partition of children). FPR stands for False Positive Rate.