DATA ANALYSIS ANTHROPOMETRIC FACIAL: A STUDY NON RANDOMIZED

Brunno Kalyxton Sousa Ramos*
Universidade de Brasília - UnB, Brasília, Brasil – bks_ramos@hotmail.com

Pedro Luiz Pinto de Lima
Universidade de Brasília - UnB, Brasília, Brasil – pedrolplb@gmail.com

Abstract

The work in question was originated from studies of a brazilian federal police agent. The objective of the agent's research was to develop and analyze effective tools to identify people through photos using the science of anthropometry.

The focus of this work was primarily to show a strategy of how to conduct an analysis of a non-randomized experiment with repeated measures in space, applied in a study of differentiation methods of anthropometric marks.

Given the small number of studies with the analysis of non-randomized trials, this work has exposed several lines of reasoning and alternatives for conducting the analysis of this type of experiment using the techniques of exploratory analysis, design of experiments, mixed models, univariate and multivariate analyzes.

It began with the exploratory analysis, observing the characteristic data through graphics and tables of averages and variances. After, we tried to create out the modeling of the experiment using the methodology of split plots and at the end of this topic, modeling with mixed models. Also, univariate and multivariate analysis was performed and, due to the type of data being analyzed and with confirmatory character, an analysis of normality of variables was also made.

The initial exploratory analysis indicated small differences between the marking procedures studied, when analyzed more deeply with mixed models, was noted that there was a statistically significant difference between them. Univariate and multivariate analyzes also confirmed this significant difference between both methods.

Keywords: repeated in space; normal test; mixed models; multivariate analysis.

1. Introduction

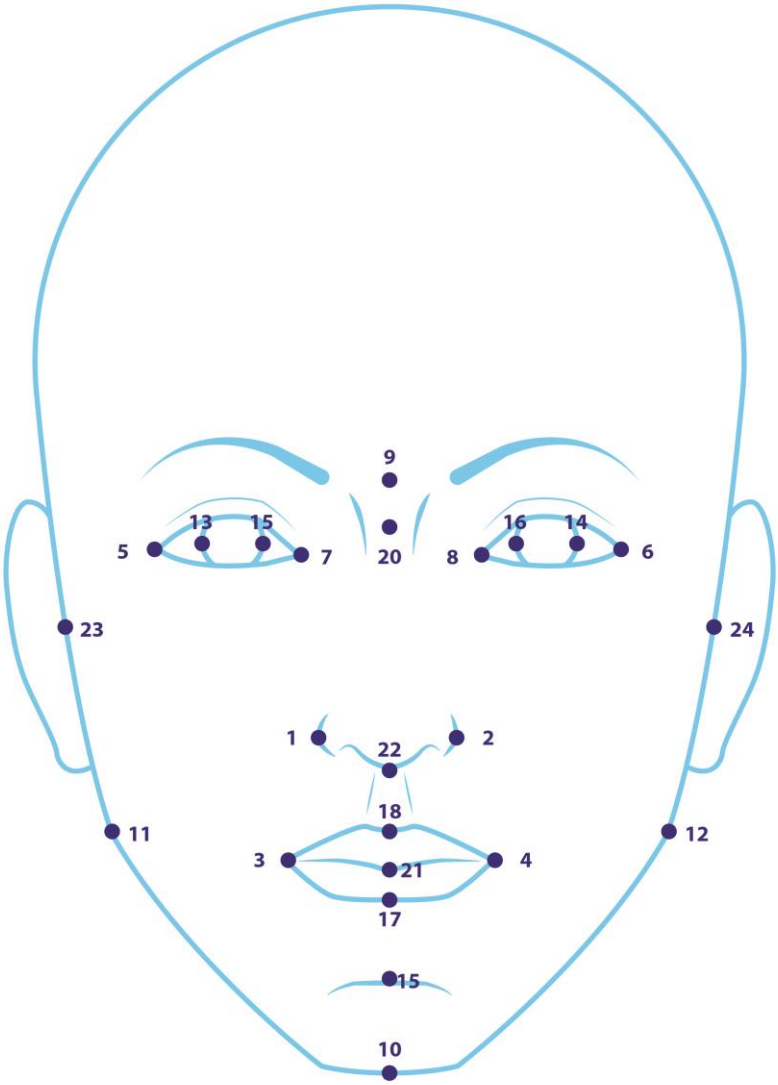
This paper aims to present an analysis of data from a comparative study between two methods of identification by facial anthropometric, which are used for the identification of suspects by the Brazilian Federal Police. This is an observational study with structured repeated in space data. The data were generated by five examiners (experts in facial identification drawing) each of which scored twenty-four facial anthropometric points that characterize the human face, from eighteen photos of eighteen distinct individuals faces, using both techniques tested (traditional and alternative). Let's briefly describe the process of origin of this data.

Anthropometry, science that studies the measures of size, weight and proportions of the human body, provides objective data in the evaluation of craniofacial morphology, through a series of measures of the head and face. In addition, it offers numerous advantages over other methods of evaluation of the morphology of the craniofacial complex by using simple techniques, non-invasive, without risk to the subject and at low cost and provide data that can be compared, since a standard was established the normal facial measurements for the Brazilian population. Differences in facial proportions are demonstrated in studies with composite populations of different races and ethnicities. In the direct anthropometry, the measure is obtained directly from the subject by means of calipers or measuring tapes. In the indirect anthropometry, the measures are collected by photography, cephalometric profile of soft tissue and computerized images of the face. All these methods have their

advantages and disadvantages and should involve four basic elements of the exam: location of anthropometric marks, measures, evaluation of findings and comparison with the normal data.

In this study will be used indirect anthropometry. For the first step, the researcher selected five examiners with experience in cephalometry and asked for that to mark the following 24 facial points: Glabella, Nasion, subnasal, Superior lip, stomion, Lower Lip, Labiomental, gnathion, Right Zygion, Left Zygion, Gonio Right, Left Gonio, Right Ectoanthion, Left Ectoanthion, Right Medial Iridion, Left Medial Iridion, Right Edge Iridion, Left Edge Iridion, Right Alar, Left Alar, Right Chelion and Left Chelion.

Figure 1: Facial anthropometric points numbered in alphabetical order



A further consideration describes the study as a non-randomized design with repeated data in space and in this case, multiple steps are required for data analysis. At first we applied the basic techniques such as analysis of graphs and exploratory statistics, which are summarized here in means,

variances and standard deviations of each anthropometric point analyzed, so using the graphical support to better visualize the data configuration. As a guide to analyze the data, we follow largely the book Hoaglin, Mosteller and Tukey (1991) which contains best practices for analyzing this type of data.

2 Exploratory Analysis

In statistics, an exploratory analysis of structured data is the submission of basic measures such which means, variances and standards deviations. For the calculation of these descriptive measures, the values of coordinates X and Y were divided by 10 in order to facilitate viewing of the obtained values.

In addition to the most common graphics, as the scatterplot, another type of graph to exploratory analysis was proposed by Andrews (1972), suitable for multivariate data. If $X_1, X_2, ..., X_p$ are the variables observed in an element of a set, function is defined

$$f(t) = X_1/\sqrt{2} +, X_2 sen(t) + X_3 \cos(t) + X_4 sen(2t) + X_5 \cos(2t) + ...$$

and shapes a graph $(f(t), t)$ with $-\pi < t < \pi$, for each data element. The idea is to represent each element that is in a multi-dimensional space, with a curve in a two-dimensional space.

In this case, the individual is characterized by 24 variables (sites) for the Y dimension and the other 24 for the X dimension, and the Y dimension to function is of the form

$$f(t) = Y_1/\sqrt{2} +, Y_2 sen(t) + Y_3 \cos(t) + Y_4 sen(2t) + Y_5 \cos(2t) + ...$$
$$Y_{22} sen(11t) + Y_{23} \cos(11t) + Y_{24} sen(12t)$$

As the study aims to discriminate differences between methods, we can get the means for the 24 sites for each individual in each method, based on 5 examiners and then get the Andrews' function for each individual and method.

Following the recommendations of Hoaglin, Mosteller and Tukey (1991), which the ANOVA can be applied to any set of structured data, we will do a variance analysis for the X and Y coordinates for each location.

3. Design of Experiments

Experiments have been made to assess and compare treatments. A lot of cares in the experiments were already taken into account before and during the experiment, but it was Fisher who introduced the randomization, an action that takes out from the researcher de decision to decide which parts will receive a particular treatment. Fisher applied a probabilistic model to indicate which portions would receive treatment. Although he wrote on the subject from 1922, was in his book, *The Design of Experiments*, in 1935, that randomization was actually placed in evidence. Subsequently, several authors studied the randomization proposed by Fisher and explored the properties that it could come. Kempthorne (1952) was one of the scientists, but others such as Nelder (1965) also contributed greatly to the theory of randomization in experiments.

One consequence of randomization is that it defines the statistical model of the experiment, i.e., how randomization was performed that defines the model, randomized blocks, divided portions, etc. It also defines how to analyze the experiment by ANOVA technique, also introduced by Fisher. Although Fisher (1935) had commented, Kempthorne and his followers concluded that the randomization its validating the estimates and significance tests in ANOVA, not normality and independence of errors. They also concluded that the randomization that defines the structure of variance and covariance of the experimental data.

Given this knowledge it is a fact that we do not know the experimental model that is being analyzed and the corresponding structure or variance and covariance since there was no randomization

in the basic experiment involving individuals, examiners and methods. Even here Fisher's ANOVA can be used, but with some care. First we will work with possible models for this experiment.

The simplest model would be a randomized block design with split plots, 5 x 2 factorial in these plots and l spots in sub plots. For the Y coordinate is the following:

$$Y_{ijkl} = \mu + E_i + M_j + (EM)_{ij} + I_k + \varepsilon_{ijk} + L_l + (EL)_{il} + (ML)_{jl} + (EML)_{ijl} + \phi_{ijkl}$$

E_i = effect of i examiner.

M_j = effect of j method.

I_k = effect of k individual.

ε_{ijk} = experimental error.

L_l = effect of l spot.

$(EM)_{ij}, (EL)_{il}, (ML)_{jl}$ and $(EML)_{ijl}$ = effects of their respective interactions.

ϕ_{ijkl} = error generated by sites

The model would be the same for X coordinate.

The second model (for Y and X) is the same model, but using the theory Huynh and Feldt (1970) and explained by example by Milliken and Johnson (2009), which proposed to adjust the degrees of freedom for the sources of variation that involve spots, since they were not randomized sequence. This adjustment was previously proposed by Greenhouse Geisser and (1959), but the former is better known. Basically, Huynh and Feldt (1970) established the conditions under which experiments with data repeated in time and space or could be analyzed as if they were sub plots and had been randomized. The adjustment of degrees of freedom is done by estimating the parameter θ of Box, has the following limits:

$$\frac{1}{t} < \theta < 1$$

With *t* being the number of spots, if θ is close to 1 then we say that there is a HF otherwise, did not. In this case, the degree of freedom for each source of variation is multiplied by θ the analysis moves on.

The third model is the same as the first but as a mixed model in which the structure of the local variance and covariance is not known. Considering the distances between both methods involving the two coordinates, we would also have those three models mentioned above: split plot, adjusted degree of freedom split plot and a mixed model. The basic model is the following:

$$D_{ijk} = \mu + E_i + I_j + \varepsilon_{ij} + L_k + (EL)_{ik} + \phi_{ijk}$$

4. Multivariate Analysis

An experiment with repeated in space data, like what is being analyzed, can be studied under the following point of view. The experiment consists of eighteen individuals, which will be blocks, five examiners and two methods are the treatments in the form of a factorial 5 x 2. It can be considered the observations at each site as a variable response of the experiment and therefore 24 variables answer to the Y coordinate, another 24 for X and 24 for distances. From this perspective, the analysis model becomes multivariate and in this case, with multivariate analysis of variance.

The multivariate model is constructed by attaching 24 univariate models in one. Using a matrix form, we define the vector $\mathbf{Y}_l$ as the vector with the observations of l site, $\mathbf{X}$ as the incidence matrix of treatments, average and blocks, β_l the vector containing the effects of treatments, effects of

the blocks and the average for the local l and ε_l the vector of errors of location l. With this notation the multivariate model is represented as follows:

$$[\mathbf{Y_1}\ \mathbf{Y_2},....,\mathbf{Y_{24}}]=[\mathbf{X}\beta_1\ \mathbf{X}\beta_2,....,\mathbf{X}\beta_{24}]+[\varepsilon_1\ \varepsilon_2,...,\varepsilon_{24}] \text{ or } \mathbf{Y}=\mathbf{X}\beta+\varepsilon$$

Being $\mathbf{Y}$ a matrix 180x24, $\mathbf{X}$ a matrix 180x29, β a matrix and 29x24 and ε a matrix 180x24. Also matrices $\mathbf{X}$ and β can be partitioned to separately represent the means, treatment's effects and the effects of blocks:

$$\mathbf{Y}=\mathbf{1}\mu+\mathbf{X}_t\mathbf{t}+\mathbf{X}_b\mathbf{b}+\varepsilon$$

or even more to separate examiners ($\mathbf{t}_1$), methods ($\mathbf{t}_2$) and interaction ($\mathbf{t}_3$) and their respective matrices:

$$Y = 1\mu + X_1t_1+X_2t_2 + X_3t_3 + X_b b + \varepsilon$$

All components of this model are matrices with their respective dimensions.

As for the inference of multivariate model the first inference is about β for the model $\mathbf{Y}=\mathbf{X}\beta+\varepsilon$. The estimate of β is made by individual estimates of minimum squares of each component of β. In the case, 24 betas can be used in the inference for each individual site. Explicitly

$$\hat{\beta}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

The second inference is about tests of hypotheses, that in the univariate ANOVA is written the way $H_0\colon \mathbf{C}\beta=\mathbf{0}$, $H_a\colon \mathbf{C}\beta\neq\mathbf{0}$ and the F test is used for testing them. In multivariate analysis, such hypotheses can also be tested but the general shape of the tests is as follows:

$$H_0\colon \mathbf{C}\beta\mathbf{L}=\mathbf{0}, H_a\colon \mathbf{C}\beta\mathbf{L}\neq\mathbf{0}$$

The $\mathbf{C}$ matrix defines the test between parameters of each location separately, and the $\mathbf{L}$ matrix defines the tests between sites that are the main gain of multivariate analysis. As an example, suppose we want to test the difference between the two methods for each site, i.e., method 1 is the same as method 2 in all locations; so we'll use $H_0\colon \mathbf{C}\beta=\mathbf{0}$, $H_a\colon \mathbf{C}\beta\neq\mathbf{0}$ for a specified $\mathbf{C}$ matrix. On the other hand, if the hypothesis is that the difference between methods is one for site 1 and another for site 2 (interaction between locals and methods), then the hypothesis is of the form $H_0\colon \mathbf{C}\beta\mathbf{L}=\mathbf{0}$, $H_a\colon \mathbf{C}\beta\mathbf{L}\neq\mathbf{0}$, to a certain $\mathbf{L}$ matrix. Also, as well as in the univariate analysis there are two basic statistics for a multivariate test.

Suppose the test is of the form $H_0\colon \mathbf{C}\beta\mathbf{L}=\mathbf{0}$, $H_a\colon \mathbf{C}\beta\mathbf{L}\neq\mathbf{0}$. So the two statistics are:

$$\mathbf{H}=\mathbf{L}'(\mathbf{C}\hat{\beta}')'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}')^{-1}(\mathbf{C}\hat{\beta}')\mathbf{L}$$

$$\mathrm{E}=\mathbf{L}'(\mathbf{Y}'\mathbf{Y}-\hat{\beta}'(\mathbf{X}'\mathbf{X})\hat{\beta}')\mathbf{L}$$

similar to those of the univariate analysis.

The most common multivariate test are the following:

Wilks' lambda = $\dfrac{\det(\mathbf{E})}{\det(\mathbf{H}+\mathbf{E})}$

Pillai's trace = $trace(\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1})$

Hotelling-Lawley's Trace = $trace(\mathbf{E}^{-1}\mathbf{H})$

Roy's Test = highest eigenvalue $(\mathbf{E}^{-1}\mathbf{H})$

For the experiment that is being analyzed, for each source of variation we have matrices **C, L, H and E** used in the 4 tests.

5. Conclusions

The graphics that were made by individual analyzes of coordinates and distances and the multivariate analysis of the coordinates and distances between methods, we come to converging conclusions: There is difference between the examiners, between methods and strong evidence of interaction between examiners and methods. Regarding locals, we also conclude that there is strong interaction between locals and examiners and between locals and methods. On the other hand, the sample is large and we can use the central limit theorem for multivariate and univariate data, reported by Vonesh and Chinchilli (1997) and consider valid inferences from the analysis.

Another important point concerning the conclusion is the discussion on what the researcher expected. It was expected that there were differences between the methods and interactions? Yes, it was expected, as was expected to find differences between the methodologies, in particular the methodology 2 (alternative) would have a smaller variance and so may be a good method to use, especially when used in discontinuous draws that is when there is the possibility of exchanging the examiner during the process. Here we paraphrase Miller (1981): the significance in a study should not be only statistical, also have to be significant in the social, economic, physiological...

References

GNANADESIKAN, R. (1977) **Methods for Statistical Data Analysis of Multivariate Observations.** Wiley.

HOAGLIN, D. C. ; MOSTELLER, F. ; TUKEY, J. W.(1991). **Fundamentals of Exploratory Analysis of Variance.** Wiley.

JOE, H. (1997) **Multivariate models and dependence concepts.** ed. Chappman & Hall/crcKUTNER, M. H. **Applied linear statistical models.-** Rev. ed. de: Applied linear regression models, quarta edição, 2004.

LITTELL, Ramon C.; MILLIKEN, George A.; STROUP, Walter W.; WOLFINGER, Russell D.; SCHABENBERGER, Oliver. (2006). **SAS® for Mixed Models.** Second Edition. Cary, NC: SAS Institute Inc.

MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. (1979). **Multivariate Analysis.** Academic Press.

MILLIKEN, George A.; JOHNSON, Dallas E. (1989). **Analysis of messy data.** 2 v. : ill.

MONTGOMERY, Douglas G. (2001). **Design and Analysis of Experiment.** Fifth edition, John Wiley & Sons, INC.

PEREIRA, Cimar A. (2013). **Manual de antropometria.** Pesquisa nacional de saude. IBGE, Rio de Janeiro.

PREEDY, Victor R. (2012). **Handbook of antropometry**, physical mesures in human form in health and disease, volume 1, parts 1-6. Ed Springer. King`s college London.

SILVA, Giovani L. (2000). **Modelos Lineares Generalizados - da teoria à pratica.** -M. Antonia Amaral Turkman DEIO/FC e CEAUL, Universidade de Lisboa E DM/IST e CMA, Universidade Técnica de Lisboa – Lisboa.

TINSLEY, H.; BROWN, S. (2000). **Handbook of Applied Multivariate Statistics and Mathematical Modeling.** Academic Press.

VIVALDI, Lucio J. (1999). **Análise de experimentos com dados repetidos ao longo do tempo ou espaço.** EMBRAPA DOCUMENTOS n.8 p1-52.

VONESH, E.F.; CHINCHILLI, V.M. (1997). **Linear and Nonlinear Models for the Analysis of Repeated Measurements.** CRC Press.