e-PS, 2012, **9**, 17-22 ISSN: 1581-9280 web edition ISSN: 1854-3928 print edition

www.Morana-rtd.com © by M O R A N A RTD d.o.o.





published by

DATA MINING ANALYSIS OVER A DATABASE OF MATERIALS OF PORTUGUESE PAINTINGS FROM THE 15TH AND 16TH CENTURIES

Vanessa H. Antunes^{1,*}, Orlando Anunciação^{2,*}, António Candeias³, Ana Isabel Seruya⁴

This paper is based on a presentation at the ICOM-CC Scientific Research Working Group Interim Meeting in Pisa, Italy, 7-8 October 2010.

Guest editor: Dr. Ilaria Bonaduce

TECHNICAL PAPER

- 1. Instituto História da Arte da Faculdade de Letras da Universidade de Lisboa (IHA-FLUL), Laboratório de Conservação e Restauro José de Figueiredo do Instituto dos Museus e da Conservação (LCRJF-IMC), Lisbon, Portugal 2. Instituto Superior Técnico, Technical University of Lisbon, INESC-ID, Lisbon, Portugal
- 3. Laboratório de Conservação e Restauro José de Figueiredo do Instituto dos Museus e da Conservação (LCRJF-IMC), Lisbon, Portugal
- 4. Centro de Física Atómica da Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal
- * joint first authors and corresponding authors: vanessahantunes@gmail.com orlando@kdbio.inesc-id.pt

received: 15.10.2010 accepted: 21.11.2011

key words:

pigments, ground layer, attribution, decision tree learning, Fisher's exact test,

This paper explores laboratory analyses of ground layers of Portuguese wooden painted panels of the 15th and 16th centuries, performed in the Laboratório de Conservação e Restauro José de Figueiredo – Instituto dos Museus e da Conservação. Based on this information, a database of materials was built. A subset of the database was selected to perform a data mining analysis.

We used a method based on decision tree learning, Fisher's exact test and permutation testing and it was possible to find a ground layer technique that distinguishes António Nogueira's Ferreira do Alentejo retable from other paintings in our database. This constitutes a small contribution to a better understanding of Portuguese painting since it was possible to establish that António Nogueira used coloured ground layers, which was an emerging technique at the time. Another contribution of this work is to show that the methodology we present here can be applied to other case studies based on similar data.

1 Introduction

Data mining is the process of extracting hidden knowledge from data. It consists of three main steps:

- pre-processing (or preparation) of the datal,
- application of data mining algorithms^{II},
- validation of results^{III}.

Data mining techniques can discover patterns in data explaining observed differences. A model is induced from the data to explain the variable of interest. Decision tree learning is a data mining algorithm that is able to produce interpretable models that can be validated with domain knowl-

I. The level of consistency of the laboratory data is the result of years of routine analysis and standard scientific investigations using different analytical methods in the IMC laboratory.

II. In bioinformatics, permutation tests have become a widely used technique. The reason for this popularity has to do with its non-parametric nature, since in many bioinformatics applications there is no solid evidence or sufficient data to assume a particular model for the obtained measurements of the biological events under investigation).19

III. The measure used in the original methodology for breast cancer was Fisher Exact Test and not the logarithm of FET. The reason for this is that we had to deal with very small numbers and it was therefore easier to work with logarithmic values. For the permutation testing procedure it is the same to use FET or the logarithm of FET because the logarithm is a monotonic function and we only need to be able to compare values for different iterations.

edge. This kind of data mining analysis has the potential of providing new insights about differences in the techniques used to produce the paintings.

The laboratory analytical data (made available by the Laboratório de Conservação e Restauro José de Figueiredo do Instituto dos Museus e da Conservação, LCRJF-IMC) and general data about each painting were collected into a single database. One of the major advantages of this was to enable the analysis of the data in many different ways, including data mining as discussed in this paper. The goal is not to focus excessively on the results, but to show that the methodology presented can be usefully applied to the kind of data.

The methodology applied in this study consisted on automatically building a decision tree model that used information about stratigraphic ground layers to discriminate between painters. We focused on the ground layers of the paintings of a single retable dated from before 1565 painted by the regional painter António Noqueira. This painter has been studied recently, resulting in a publication of analytical data on 6 paintings (of a total of 8) from the "Ferreira do Alentejo Retable". 1The question posed during data mining analysis was whether the painter used a special ground layer technique that distinguishes him from other painters in the database. The results showed that António Noqueira used a pattern, based on the presence or absence of pigments in the ground layers, that differentiates him from earlier painters in our database.

The presented work is part of a project with the objective to perform a study of ground layer techniques used by Portuguese painters and workshops from the end of gothic and renaissance to mannerism. We therefore needed to explore the relationships between painters and materials used in the creation of their works.

2 Database Description

Before using the database, we consulted information used in other existing painting databases such as NARCISSE (Network of Art Research Computer Image Systems in Europe),² JOCONDE (catalogue collectif des collections des musées de France)³ or the Raphael Research Resource project,⁴ developed

by the National Gallery of London and its partners. Portuguese models consulted were the IMC's database MATRIZPIX⁵ and the database developed for the study of the Portuguese painter Gregório Lopes and his workshop.⁶

The definition of attributes necessary to unequivocally identify paintings, their materials and the 'thesaurus' to be used, followed the research, specifications of standards, and specific inventory of works of art at the LCRJF- IMC. Each painting has associated identification and art historical data: title, author, workshop, date, material, technique, dimensions, inscriptions, owner Institution, observations, identification number of the laboratory process and its image. In the registration of the stratigraphic constituent materials of paintings, data systematisation used in IMC's microscopic analysis worksheets was followed.

For each sample we saved the following: identification of the laboratory number, general colour, location and number of layers. The database also included data on each layer found in the corresponding stratigraphy: layer number, thickness, grain colour, pigments and binding media, recording chart on the location of the sample in painting and photographic record of the stratigraphy.

The relational scheme with the most relevant tables of the database for this work is presented in Fig. 1. Tables for 'Painting', 'Stratigraphy', 'Layer', 'LayerPigment', 'Pigment', 'PaintingAuthor' and 'Author' can be seen.

Currently, the database contains the information of about 90 paintings of the 15th and 16th centuries. The LCRJF-IMC laboratory analytical data include a total of 3863 layers belonging to stratigraphies, 1033 of which are ground layers.

3 Methods

In this section we present a brief description of the methods used in this work.

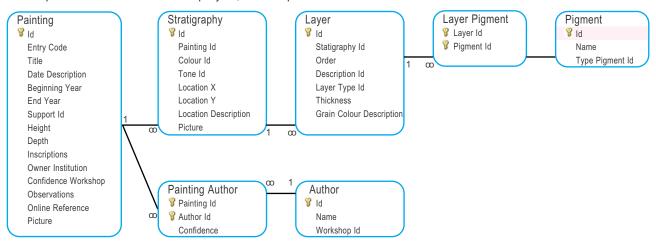


Figure 1: Relational scheme with the most relevant tables in the database.

3.1 Decision Trees

Decision Tree Learning is one of the most widely used and practical methods for classification.⁷ The algorithm produces models that are very simple to understand and interpret by domain experts such as conservators and art historians. Learned trees can be represented as a set of if-then rules that improve human readability.

A decision tree consists of nodes that have exactly one incoming edge, except the root node that has no incoming edges. A node with outgoing edges is an internal (or test) node, while the other nodes are called leaves (also known as terminal nodes or decision nodes).

Each internal node splits the instance space into two or more subspaces, according to the result of a test that considers a single attribute and its corresponding value. Each leaf is associated with one class, representing a value of the target variable given the values of the variables represented by the path from the root node. The classification of one instance is done by navigating from the root node down to a leaf, according to the outcome of the tests along the path. For this to be possible, a decision tree must cover the space of all possible instances. In a decision tree each node is labelled with the attribute it tests, and its branches are labelled with its corresponding values.

3.2 Statistical Validation

In this subsection we present the methods used to statistically validate our results.

3.2.1 Fisher's Exact Test

Fisher's exact test (FET) is a statistical test used to determine if there are non-random associations between two categorical variables.⁸

In the case of a 2x2 matrix, the conditional probability of getting the actual matrix, given the particular row and column sums, is given by Equation (1).

$$\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$
 (Eq. 1)

In Equation 1, a, b, c and d are the 4 entries of the 2x2 matrix and n=a+b+c+d. If we want to calculate the p-value of the test, we can do it by computing the sum of all p-values which are $\leq P_{\text{cutoff}}$.

To compute the value of Fisher's Exact Test given by Equation (1), we need to decompose it in order to avoid computing large factorials. An equivalent formulation of Equation (1) is given by:

$$P_{\text{cutoff}} = e^{s(a+b)+s(c+d)s(b+d)-s(a)-s(b)-s(c)-s(d)}$$
m
(Eq. 2)

in which $s(m) = \sum_{i=1}^{m} log(i)$.

This test is applied to evaluate the quality of leaves in the decision tree as described under Methodology.

3.2.2 Permutation Tests

Permutation testsⁱⁱ are non-parametric procedures for determining statistical significance based on rearrangements of the labels of a dataset. It is a robust method, but it can be computationally intensive. A test statistic, which is computed from the dataset, is compared with the distribution of permutation values. These permutation values are computed similarly to the test statistic, but, under a random rearrangement of the labels of the dataset.⁹ Permutation tests can help to reduce the multiple testing burden¹⁰ and can be used to compare statistical tests.¹¹

The main disadvantage of permutation testing is that it needs a very large number of permutations when small *p*-values are to be accurately estimated, which is computationally expensive. To address this problem, the tail of the distribution of permutation values can be approximated by a generalised Pareto distribution. According to Knijnenburg et al. accurate *p*-value estimates can be obtained with a drastically reduced number of permutations when compared with the standard empirical way of computing *p*-values.

4 Methodology

4.1 Preliminary Data Mining Analysis

Our first data mining analysis was to check if there were interesting leafs that could distinguish between pictorial centres. We ran the decision tree learning algorithm over datasets to distinguish between three main pictorial centres in our database: Lisbon, Coimbra and Évora regions. We looked at the obtained decision trees and observed that data mining analysis found patterns that could be further explored. However, in the data mining analysis of Lisbon versus Évora we found very interesting leaves concerning painting layer materials:

- Black Vegetable Charcoal is Present' and 'Gypsum is Present' and 'Lead White is Absent' → 34 layers from Évora paintings and 1 layer from a Lisbon painting.
- Black Vegetable Charcoal is Present' and 'Gypsum is Absent' and 'Yellow Ochre is Present' and 'Lead White is Present' → 31 layers from Évora paintings and 9 layers from Lisbon paintings.

In our database, for the Évora pictorial centre, we have paintings from the António Nogueira's Ferreira do Alentejo retable (6), António Nogueira's workshop (1), Francisco de Campos (1) and an Unknown Painter (1). Since most paintings of the Évora pictorial centre in our database are from António Nogueira, we decided to study this painter in more detail. For this we tried to answer the question if this painter had a specific technique that could distinguish him from other painters in our database. The following subsections detail the computational analysis focussing on ground layers.

4.2 Data Preparation

The first step in a data mining process is to prepare the data to be analysed. We focussed on preparing a dataset based on our database, which could enable us to find distinctive characteristics in terms of the pigments used by painters in the ground layers. In our database, we have a 'many-to-many' relationship between 'Layer' and 'Pigment'. In order to perform our data mining analysis, we transformed our data, building one binary variable for each pigment. That variable takes the value 1 if the corresponding pigment is present in the layer and the value 0 if the corresponding pigment is absent in the layer. At this step we have a dataset of ground layers with 46 binary variables, which correspond to the number of pigments in the computational analysis.

We also have to take into account the fact that we study painters and since one painting can have multiple painters, each instance of our dataset will represent an association between a painter and a layer that belongs to a stratigraphy of a certain painting. This means that each layer will have as many instances in our dataset as painters of the corresponding painting. Instances that did not contain information about ground layer pigments were removed from the dataset. After this filtering step we ended up with 887 instances, referring to 60 paintings from 25 different painters and 664 ground layers.

The target variable that we wish to study is the variable 'Painter'. In order to be able to analyse the differences in the ground layer preparation technique between the painter António Nogueira and other painters in our database, we created a binary variable with value 1 when the ground layer was prepared by António Nogueira and 0 when the ground layer was prepared by another painter.

The dataset was then analysed with a methodology very close to the one originally developed for the detection of high-risk breast cancer groups. 12

4.3 Application of Data Mining Algorithms

We applied decision tree learning to our dataset with the goal of predicting our binary target variable 'Painter'. To conduct our experiments, we used the Weka 3.6.0 J48 Decision Tree which generates a C4.5 decision tree. Weka is a widely used open source data mining package (University of Waikato). 13

After the model was built, we looked into the tree, selected the best leaf L according to the logarithm of Fisher's Exact Test (FET) valueⁱⁱⁱ (lower values of logFET mean stronger associations) and saved this value, $\log FET(L)$. The best leaf L corresponds to a specific combination of the presence and absence of pigments in the ground layer that possibly distinguishes António Nogueira from other painters. In order to check if this observation has statistical meaning, we need to statistically validate it with permutation tests.

4.4 Statistical Validation

In order to statistically validate our detected ground layer technique, we used 10,000 permutation tests. This number of permutation tests was chosen since it enables us to estimate a p-value as low as 10-4, which we consider to be highly statistically significant. We state as our null hypothesis H_0 that there is no association between our pigment variables and target variable 'Painter'. IV If we find a strong association between one or more pigment variables and the variable 'Painter', then we can reject the null hypothesis. In order to test whether the null hypothesis can be rejected, for each permutation test we randomly permute labels (target variable 'Painter'). We then build a new decision tree with the same parameters as in our original dataset. We check all leaves from the decision tree and save the lowest logFET value. This means that in each permutation test we save the logFET value of the best leaf. In the end of the execution of the permutation tests, we have 10,000 logFET values that were generated according to the underlying and unknown distribution of the null hypothesis. We can then check where the logFET value of our leaf L stands on the estimated distribution of our null hypothesis. This way we can see if logFET(L) is extreme enough so that we can reject the null hypothesis with high statistical confidence, therefore obtaining our adjusted p-value.

5 Results and Discussion

In Fig. 2, the decision tree automatically learned over the entire dataset is presented. We can see that the attribute on the root of the tree is 'Black Vegetable Charcoal' which was identified by the Decision Tree C4.5 algorithm as being the most important attribute for the distinction between António Nogueira's ground layer technique and other painters in the database. There is a very interesting leaf L in this tree: 'Black Vegetable Charcoal is Present' and 'Lead White is Absent'. We have 36 instances in our dataset that have this characteristic, 34 of which are ground layers that belong to António Nogueira's Ferreira do Alentejo retable. The other two instances correspond to a ground layer in a painting of Simão Rodrigues and Domingos Vieira Serrão. V, 14. This interesting leaf L has a value of logFET(L)= -110.95. This value is a measure of how good our leaf L is.

If António Nogueira did not have any unique technique when compared with the other painters, we would expect to obtain similar values for this measure applied to the best leaf in each decision tree of the permutation testing procedure. However, the results do not show a permutation test with a lower logFET value than in our original dataset. The adjusted *p*-value obtained with 10,000 permutation tests as described in section 4.4 is therefore lower than 10⁻⁴. With these results we can reject our null hypothesis at the 0.01% level of confidence.

The distinguished ground layers are part of one subgroup with no lead white added. In these coloured ground layers there is a mixture of gypsum and black

IV. The null hypothesis can be also stated as: António Nogueira did not have any unique technique related to the use of pigments in the ground layer when compared with other painters in the database.

V. We have to take into account, however, that these two painters, Simão Rodrigues and Domingos Vieira Serrão are painters from the beginning of the 17th century.

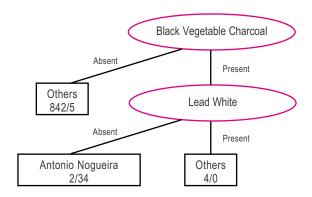


Figure 2: Decision Tree Model learned on the dataset. In the leaves there is information about Other Painters/António Nogueira distribution. In the first leaf, when 'Black Vegetable Charcoal is Absent', we have 842 associations between layers and other painters and 5 from António Nogueira. In the second leaf, when 'Black Vegetable Charcoal is Present' and 'Lead White is Absent', we have 2 associations between layers and other painters and 34 layers from António Nogueira. In the third leaf, when both 'Black Vegetable Charcoal is Present' and 'Lead White is Present', we have 4 associations between layers and other painters and none from António Nogueira.

vegetable charcoal. In addition in some of these layers there is also yellow ochre.

In the 15th and the beginning of 16th century, the most common ground layer in Southern Europe is white gypsum, ¹⁵⁻²⁰ however, it has to be taken into account that the tempera and oil retable in questions is dated to around 1565.¹ At that time, in Italy, coloured surfaces in preparatory layers (grounds and primings) were already subject to significant changes with respect to the painting technique,²¹ despite further use of white gypsum well into the 17th century.²²⁻²³

This fact could contribute to António Nogueira being distinguished as a vanguard painter not only in terms of his early mannerist iconographic style (as referred to by Joaquim Oliveira Caetano^{1,24-26} and Vitor Serrão²⁶⁻²⁷), but also in the use of ground layer techniques in one of his final works, the Ferreira do Alentejo retable.

6 Conclusions

The results obtained shed further light on the Portuguese painter António Nogueira. We were able to find a ground layer technique that distinguishes him from most of the other painters in our database. This represents a small contribution to better understanding of Portuguese painting since it was possible to establish that Nogueira used coloured ground layers, which was an emerging technique at the time of the execution of Ferreira do Alentejo Retable. However, in order to have more confidence in the results, more paintings in the same time frame of Ferreira do Alentejo Retable would be needed in the database.

Another contribution of this work is to show that the methodology we presented here can be applied to similar studies on this kind of data. It is our current plan to study other painters and workshops using the methodology, and also use other data mining techniques. A big advantage of this methodology is that

we let the decision tree learning algorithm discover the patterns that better distinguish the painters involved. The use of permutation tests enables us to have a degree of statistical confidence about our results. In this way, we obtained a data mining algorithm discovering patterns that are easily humanly interpretable and are still able to provide some statistical confidence of the differences observed.

7 Acknowledgments

The authors wish to acknowledge the Fundação para a Ciência e Tecnologia for financial support (PhD grants SFRH / BD / 37929 / 2007 and SFRH / BD / 41984 / 2007 and project "The invisible ground layer and its influence in Portuguese paintings of the 15th and 16th centuries: a question to be settled" - PTDC/EAT-HAT/100868/2008) through the programme Ciencia e Inovação 2010 (POCI 2010).

Vanessa H. Antunes thanks to Drs. Vítor Serrão, Maria Isabel Ribeiro, and João Coroado, to Instituto de História da Arte da Faculdade de Letras da Universidade de Lisboa and Laboratório de Conservação e Restauro José de Figueiredo do Instituto dos Museus e da Conservação. Orlando Anunciação thanks to Prof. Arlindo Oliveira and Prof. Susana Vinga for the support in the development of the methodology applied in this work.

8 References

- Retábulo de Ferreira do Alentejo, Projecto "Estudos e Investigação sobre o Património Cultural" Divisão de Documentação e Divulgação, Instituto Português de Conservação e Restauro, Lisboa, 2004.
- 2. NARCISSE: Network of Art Research Computer Image SystemS in Europe, Sistema documental de pintura e iluminura, Arquivos Nacionais, Torre do Tombo, Lisboa, 1993.
- 3. JOCONDE :catalogue collectif des collections des musées de France.

http://www.culture.gouv.fr/documentation/joconde/fr/apropos/presentation-joconde.htm#historique (accessed 15/10/2010).

- 4. Raphael Research Resource project, N.G.o. Art,http://cima.ng-london.org.uk/scientific/index.php, 2010 (accessed 15/10/2010).
- 5. MatrizPix:colecções da Rede Portuguesa de Museus do IMC, http://www.matriznet.imc-ip.pt (accessed 15/10/2010).
- Estudo da pintura portuguesa, Oficina de Gregório Lopes: Actas do Seminário Internacional, Instituto de José de Figueiredo, Lisboa, 1999.
- 7. T. M. Mitchell, Machine Learning, McGraw-Hill, 1997.
- 8. E.W. Weisstein, *Fisher's exact test*, From MathWorld A Wolfram Web Resource, http://mathworld. wolfram.com/FishersExactTest. html (accessed 15/10/2010)..
- 9. T.A. Knijnenburg, L.F.A. Wessels, M.J.T. Reinders, I. Shmulevich, *Fewer permutations, more accurate P-values*, Bioinformatics 2009, **25**, 161.
- 10. J. Marchini, P. Donnelly, L.R. Cardon, *Genome-wide strategies* for detecting multiple loci that influence complex diseases, Nature qenetics 2005, **37**, 413-417.

- 11. M. Li, K. Wang, S.F.A. Grant, H. Hakonarson, C. Li, *ATOM: a powerful gene-based association test by combining optimally weighted markers*, Bioinformatics 2009, **25**, 497.
- 12. O. Anunciação, B. Gomes, S. Vinga, J. Gaspar, A. Oliveira, J. Rueff, *A Data Mining Approach for the Detection of High-Risk Breast Cancer Groups*, Adv. Bioinform. 43-51.
- 13. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, *The WEKA Data Mining Software: An Update*, SIGKDD Explorations, Volume 11, Issue 1, 2009.
- 14. I. Ribeiro, P. Sousa, A materialidade do Conjunto Retabular de Simão Rodrigues e Domingos Vieira Serrão, Laboratório de Conservação e Restauro José de Figueiredo do Instituto dos Museus e da Conservação, Lisboa, 2007.
- 15. L. Carlyle, J.J. Boon, R. Haswell, M. Stols-Witlox, *Historically accurate ground reconstructions for oil paintings*, in: J.H. Townsend, T. Doherty, G. Heydenreich, J. Ridge, Eds., *Preparation for Paintings: The Artist's Choice and Its Consequences*, Archetype Books, 2008, 110-122.
- 16. P. Hendy, A.S. Lucas, *The ground in pictures*, Museum 1968, **21** 245-76
- 17. S. Santos Gómez, Las preparaciones de yeso en la pintura sobre tabla de la Escuela Española, Facultad de Bellas Artes, Universidad Complutense de Madrid, Servicio de Publicaciones, Madrid, 2006, http://www.ucm.es/BUCM/tesis/bba/ucm-t28904.pdf (accessed 15/10/2010).
- 18. M. J. Gonzalez Lopez, Estudio de las preparaciones de pintura sobre soportes de tela y tabla: Caracterizacion de sus principales componentes, comportamiento y factores de deterioro, Universidad de Sevilla. Sevilla. 1992. 792.
- 19. S. S. Gómez, M. S. A. Moya, A. Rodriguez, *Reconstruction of documented preparation methods for gesso grosso and gesso sottile in Spanish School panel paintings*, in: S. Kroustallis, J. H. Townsend, E. C. Bruquetas, A. Stijnman, M. S. A. Moya, Eds., *Art Technology Sources and Methods*, Archetype Publications, London, 2008, 196.
- 20. A. Illán, R. Romero Asenjo, A. Sáenz de Tejada, Características de las preparaciones sevillanas en pintura de caballete entre 1600 y 1700: implicaciones en el campo de la restauración, Il Congreso del GEIIC. Investigacion en Conservacion y Restauracion, 9-11/11/2005, Universidad de Barcelona, 2005.
- 21. J. Dunkerton, M. Spring, *The development of painting on coloured surfaces in sixteenth-century Italy*, in: A. Roy, P. Smith, D. Bomford, Eds., *Painting techniques, history, materials and studio practice: contributions to the Dublin Congress*, 7-11/09/1998, International Institute for Conservation of Historic and Artistic Works, London, 1998, pp. 120-130.
- 22. A. J. Cruz, *Da sombra para a luz Materiais e técnicas da pintura de Bento Coelho da Silveira*, Instituto Português do Património Arquitectónico, Lisboa, 1999.
- 23. A. R. Duval, Les preparations colorees des tableaux de l'Ecole Francaise des dix-septieme et dix-huitieme siecles, Studies in Conservation, 37,1992, 239-258.
- 24. J. O.Caetano, *O antigo retábulo de pintura da capela-mor da Igreja da Misericórdia de Beja*, in: *A Cidade de Évora*, Évora no. 65-66,1982-83,197-210.
- 25. J. O.Caetano, *Novas obras do pintor António Nogueira*, in: *Anais da Real Sociedade Arqueológica Lusitana*, Santiago do Cacém, 2nd series, vol. I, 1987, pp. 77-92.
- 26. J. O.Caetano, Ao modo de Itália: a pintura portuguesa na Idade do Humanismo, in: V.Serrão, Ed., A Pintura Maneirista em

Portugal. Arte no tempo de Camões, Lisboa, Comissão Nacional para as Comemorações dos Descobrimentos Portugueses, 1995.

27. V.Serrão, A pintura maneirista em Portugal:das brandas "maneiras" ao reforço da propaganda, in: P. Pereira, Ed., História da Arte Portuguesa, vol.2, Lisboa, Círculo de Leitores, 1995,427-502.