Received: 15 December 2022 | Revised: 21 June 2023 | Accepted: 10 July 2023

DOI: 10.1002/xrs.3388

RESEARCH ARTICLE

X-RAY SPECTROMETRY WILEY

Interpretation of x-ray spectral data using self-organising maps and hierarchical clustering: A study of Vilhelm Hammershøi's use of painting materials

Gianluca Pastorelli | **Annette S. Ortiz Miranda** | **Anne Haack Christensen**

National Gallery of Denmark, Statens Museum for Kunst (SMK), Copenhagen, Denmark

Correspondence
Gianluca Pastorelli, National Gallery of Denmark, Statens Museum for Kunst (SMK), Sølvgade 48-50, 1307, Copenhagen K, Denmark.
Email: gipa@smk.dk

Funding information
Augustinus Fonden; Ny Carlsbergfondet

Abstract

The goal of the Vilhelm Hammershøi Digital Archive project of the National Gallery of Denmark is to understand the Danish painter Vilhelm Hammershøi's painting methods by compiling a comprehensive amount of data on his use of materials and working methods through visual and technical examination of a large number of his paintings, and to make this information available to researchers and the public in an open access digital resource. A clear understanding of the full suite of pigments across the paintings requires determination of which materials comprise the palettes of the ground and paint layers. Scanning electron microscopy/energy-dispersive x-ray spectroscopy and macro x-ray fluorescence spectroscopy were selected as the key analytical techniques due to their ability to yield chemical information at the elemental level. This article presents a method that combines unsupervised machine learning and cluster analysis techniques, to automatically reduce the large x-ray spectral data to sets of distinct clusters that share similar spectra, making it possible to identify materials more precisely. The proposed method allowed the grouping of materials by chemical composition, which enabled an optimal understanding of the pigments used in the ground layers sampled from a large number of paintings as well as in the paint layer examined at the surface of one selected painting. The method performed well when compared with other well-established data mining techniques, and it helped reduce the time necessary for the interpretation of the analytical results significantly. Through this approach, a basis for a more nuanced view of Hammershøi's artistic idea and technical development will be generated.

KEYWORDS

heritage science, hierarchical cluster analysis, MA-XRF, self-organising maps, SEM–EDX

1 | INTRODUCTION

The rapidly growing international interest in the work of Danish artist Vilhelm Hammershøi (1864–1916) among scholars and the public has resulted in an increased attention to the artist's working methods and technical development by museums, researchers and collectors. However, little is known about the technique and material aspects of Hammershøi's art. The 5-year Vilhelm Hammershøi Digital Archive (ViHDA) project of the

National Gallery of Denmark (Statens Museum for Kunst, SMK) aims at investigating the formation and evolution of Hammershøi's working methods, materials and techniques through technical and scientific documentation of his works.

Vilhelm Hammershøi was Denmark's internationally best-known artist around 1900.[1] Today, museums worldwide increasingly collect his paintings, and exhibitions of his works have attracted great interest. Despite a considerable body of art historical research into Hammershøi's oeuvre and iconography compiled through many years, knowledge about the technical aspect of his art such as his painting and drawing techniques, as well as his choice and use of materials, is limited. When looking at Vilhelm Hammershøi's works with their toned down colours, many will be left with the perception that the artist worked in a few shades of grey supplemented with a handful of warmer brownish colours. In reality, preliminary analyses performed at SMK as part of the ViHDA project suggested a more complex and sophisticated technique and choice of materials, with significant development throughout the artist's career. To understand Hammershøi's use of pigments, the investigation within the ViHDA project comprises visual and technical examination, imaging, scientific analyses and registration of paintings from Danish and foreign public and private collections. The project is an interdisciplinary collaboration between conservators, heritage scientists, art historians and imaging experts. It will result in an open access digital archive where data, images and other results collected during the project will be made available to scholars and the public.

The numerous advanced scientific methods used for the project include chemical imaging of paintings by macro x-ray fluorescence (MA-XRF) spectroscopy, and analysis of samples by scanning electron microscopy coupled with energy-dispersive x-ray (SEM–EDX) spectroscopy. Both techniques can map the elements in the paint and ground layers, making it possible to identify the chemical composition of Hammershøi's paints in the individual artworks and paint samples. During the first 2 years of the project, 62 paintings produced between 1884 and 1911 have been inspected by SEM–EDX and MA-XRF spectroscopies. Given the large number of x-ray spectra that are routinely measured on the samples and on the paintings, both the SEM–EDX analyses and the MA-XRF scans are resulting in an enormous amount of data that is destined to grow considerably over the course of the project. Identifying pigments and understanding their use by the artist based on x-ray spectral data requires hours of interpretation of scientific results. The standard way to analyse this data is, on the one hand, to examine individual x-ray spectra measured at specific areas, and, on the other hand, to produce and compare selected single-element maps. These methods are certainly suitable for small sets of data, while they can become challenging and can lead to incorrect or incomplete interpretations when working with very large collections of data and complex material compositions. Therefore, a process of extracting and discovering patterns in those large data sets is essential. In the field of heritage science, advanced data mining methods have been applied to x-ray spectral data before.[2–4] However, those methods often require the analyst to identify crucial parameters in advance, in order to optimise the performance of the analysis. The aim of this work was to develop and test a combination of data mining methods based on unsupervised machine learning and cluster analysis that require little to no parameter tuning. This way, the EDX and the XRF spectral data could be automatically reduced to sets of distinct clusters that share similar spectra, making it possible to identify materials more precisely and deduce the compositions of the examined ground and paint layers. The present article illustrates this approach in a twofold manner: through the examination of the EDX spectra that have been collected from 66 samples, and through the analysis of the MA-XRF spectral imaging data cube of one painting in the collection of SMK that was selected as a case study.

2 | MATERIALS AND METHODS

2.1 | Sample analysis by optical microscopy and SEM–EDX

One or two representative samples of the ground layer were collected from the edges of each of the 62 paintings, for a total of 66 samples. All samples were embedded in Technovit 2000 LC light curing resin from Kulzer Technik (Wehrheim, DE) and prepared as cross sections by polishing the transverse plane. A Leica DM2500 M optical microscope (maximum 100×) coupled with a Leica DMC4500 camera was used to examine the samples visually and to photograph the cross sections in both reflected visible light (dark field) and ultraviolet light. Afterwards, elemental analyses on the cross sections were carried out at the scientific laboratory of the Royal Danish Academy—Institute of Conservation in Copenhagen, using a Hitachi S-3400N scanning electron microscope equipped with an energy dispersive x-ray spectrometer. The spectrometer is a Bruker Quantax 200 EDX system with two Peltier-cooled XFlash silicon drift detectors (SDD), which have an active area of 20 mm^2 each. Measurements were performed in variable pressure mode

(30 Pa) on non-coated polished sections using an accelerating voltage of 20 kV, a probe current of 50 μA in backscatter mode, and a working distance of 10 mm. A combination of multipoint analysis and x-ray elemental mapping was employed. Specific areas to be examined for elemental composition with multi-point measurements were carefully selected onto SEM backscattered electron (BSE) images manually, making sure that the target areas on each layer were representative of the entire layer. Large particles of single pigments were measured individually for pigment characterisation purposes, and were not included in the areas representative of the layers. The x-ray elemental mapping was used to visualise the distributions of the elements present in each layer. The acquisition times (live time) for analysing each selected area and for producing the elemental maps were 60 and 600 s, respectively.

2.2 | Painting analysis by MA-XRF

The painting selected as a case study for this report is a high format (43.5 cm × 27.5 cm) portrait of the artist *Jens Ferdinand Willumsen* (1901, oil on cardboard). The portrayed artist is dressed in a dark jacket with a white shirt and a dark bowtie, and gazes directly at the viewer with a slight leftward turn of the head, casting a shadow from the prominent nose (Figure 1). MA-XRF elemental mapping was performed at the front side of the painting in the laboratory for Conservation and Art Technological Studies (CATS) of the department of Conservation and Scientific Research (BENA) at SMK using a Bruker CRONO system developed by XGLab S.R.L. The instrument consists of a measuring head with an Rh-target microfocus x-ray tube (10 W, maximum voltage 50 kV, maximum current 0.2 mA), and a 50 mm^2 SDD with beryllium window (energy resolution <140 eV at Mn Kα). The measuring spot can be varied by changing the distance between the paint surface and the measuring head. The instrument was operated at 50 kV and 60 μA. The elemental two-dimensional (2D) mapping of the painting's surface was achieved through an automatic XY-motorised stage with a 30 ms/pixel acquisition time and a 1 mm step size in both the horizontal and vertical directions. The acquired data cube was processed using open-source PyMCA software (version 5.6.7).

2.3 | Data mining

2.3.1 | Data pre-treatment

All EDX data was recorded in a single data matrix composed of 66 observations corresponding to the measured

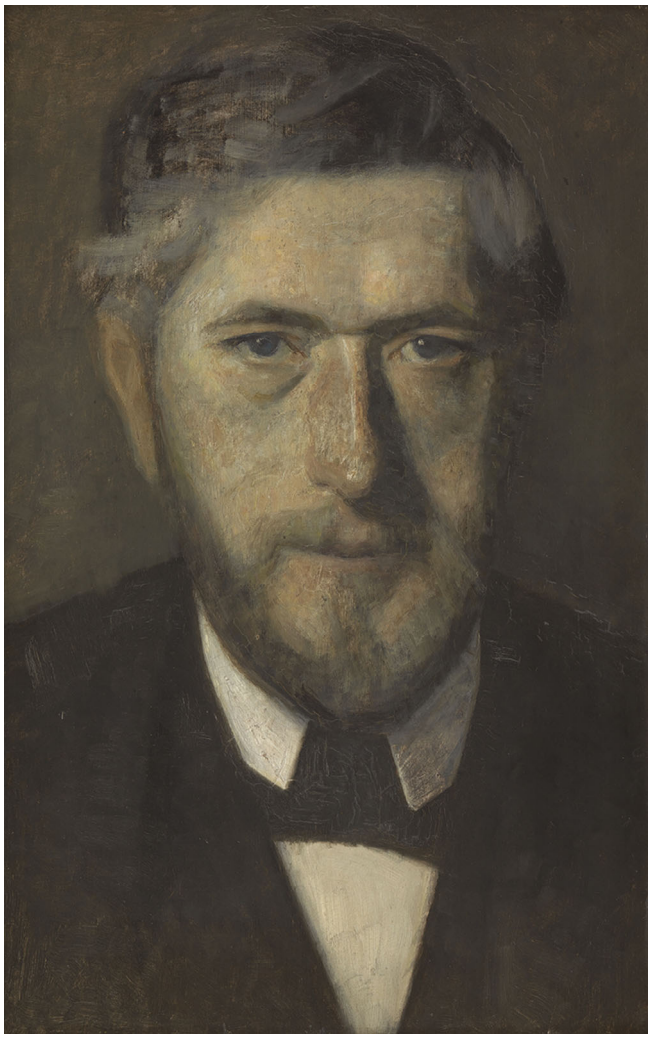


FIGURE 1 Vilhelm Hammershøi. *Jens Ferdinand Willumsen, Study for Five Portraits*. 1901. Oil on cardboard, 43 cm × 27.5 cm. KMS6793, National Gallery of Denmark (SMK), Copenhagen. [Colour figure can be viewed at wileyonlinelibrary.com]

samples, and 3860 variables associated with the x-ray energies (in the 0.7–20.0 keV range) of the sum spectrum of each sample. Because the morphology of the ground layers varies significantly across different samples, the net intensities for each variable were normalised (i.e., centred and scaled to mean 0 and standard deviation 1) within their respective ranges.

Pre-treatment of the XRF data cube involved fitting of the spectra. This was necessary not only for reducing the amount of data and to improve processing efficiency, but also for removing noise-dominated signals of x-ray energies associated with absent elements or signals of unwanted emission lines (e.g., those attributed to elements present in the ground layer), so as to narrow down the clustering result to the pigments in the paint layer. XRF data was recorded in a single data matrix composed of 116,724 observations corresponding to the measured points on the painting surface, and 14 variables associated with the net intensities of representative elements

clearly distinguishable in terms of energy emissions, specifically: P K, Pb M, Cd L, K K, Ca K, Ba L, Ti K, Cr K, Mn K, Fe K, Co K, Ni K, Cu K and Hg L. The selected variables are assumed to contain significant information about the pigments present in the paint layer, as confirmed by the respective elemental maps (Figure 2). Some other spectral lines were discarded, namely Pb L and Zn K, due to their association with the ground layer that was already investigated by SEM–EDX, thus separating their signals from those of the other elements present in the paint layer. No normalisation of the intensity values was performed in the XRF data set. Both data sets were then analysed using a combination of two approaches: unsupervised machine learning by self-organising maps (SOM) and cluster analysis by hierarchical clustering (HCA). Both techniques were performed in the R (version 4.2.2) environment using the kohonen (version 3.0.11),[5,6] aweSOM (version 1.3)[7] and maptree (version 1.4-8)[8] packages on a standard mid-range laptop PC. An R file containing the script for the SOM-HCA process can

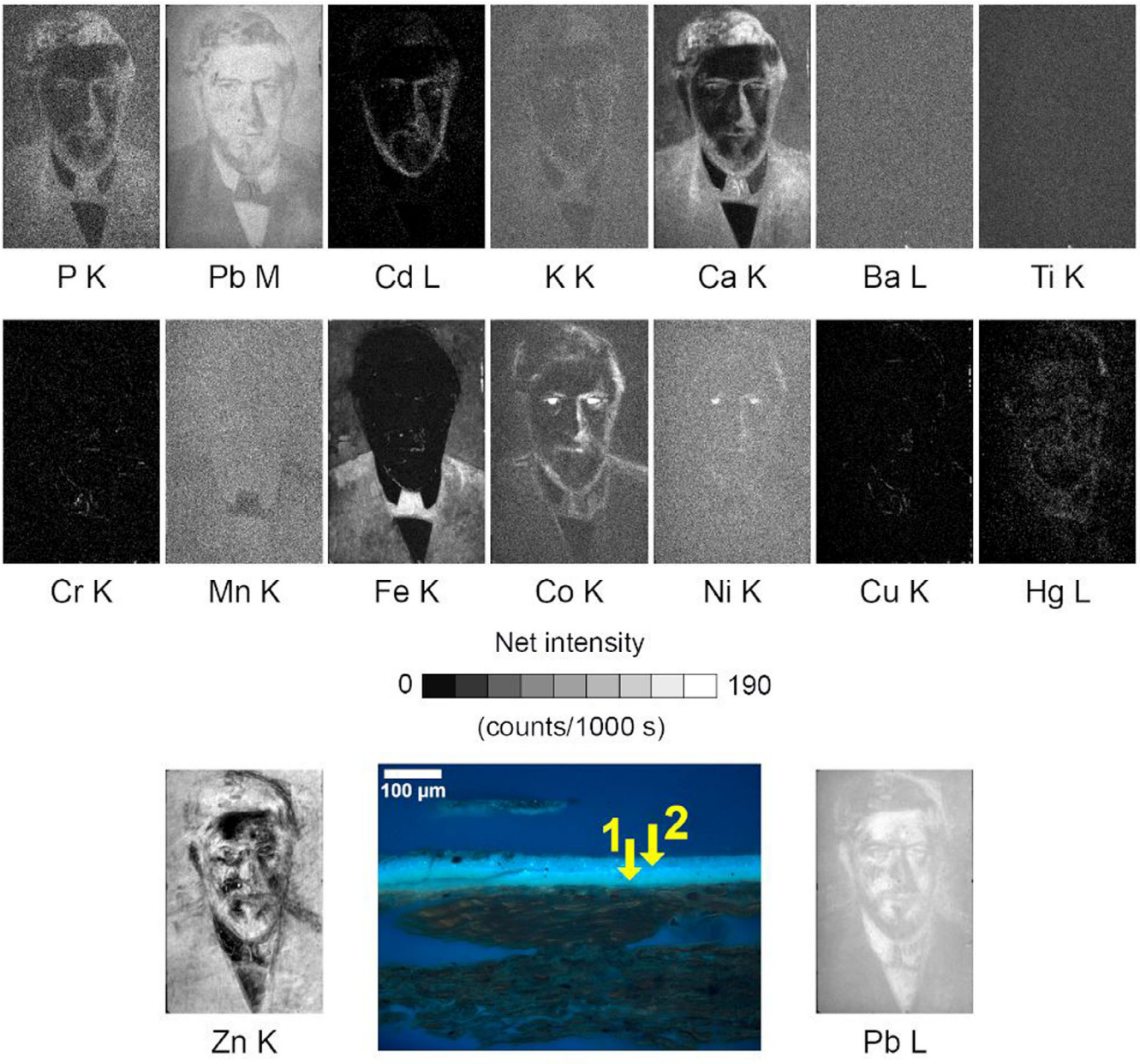


FIGURE 2 Elemental distribution maps of the portrait of *Jens Ferdinand Willumsen*. The bottom row displays the elemental distribution maps of the Zn- and Pb-based double ground layer alongside a cross section taken from the painting and imaged using optical microscopy under UV radiation. The cross section shows the ground preparation composed of a lead white application (layer 1) and a subsequent zinc white application (layer 2). [Colour figure can be viewed at wileyonlinelibrary.com]

be accessed freely from this GitHub repository: https://github.com/Gianluca-Pastorelli/SOM-HCA.git.

2.3.2 | Unsupervised machine learning by SOM

To increase the interpretability of each data set while preserving the maximum amount of information, and to enable the visualisation of their main characteristics, SOM, a neural network-based unsupervised data visualisation technique used to display high-dimensional data sets in 2D representations,[9] was performed on both the EDX and XRF data sets. The distinctive aspect of SOM is that the topological features of the original input data are preserved on the output map, meaning that similar observations are placed close together on the resulting SOM grid. For instance, all samples containing elements characteristic of earth pigments (e.g., Al, Si and Fe) will be mapped to groups (called nodes or neurons) in the same area of the grid. In general, the first step is to specify the size of the training grid before training the SOM. This can be achieved by utilising a heuristic formula proposed by Vesanto et al.,[10] which is currently used in the *som_topol_struct()* function of Matlab's SOM Toolbox and allows calculating the optimal number of nodes using the following Equation (1):

$$N = 5 \times \sqrt{n}, \tag{1}$$

where N is the number of nodes in the SOM grid and n is the number of observations in the data set. However, this approach often leads to the creation of unnecessarily large maps with a high number of empty nodes. Although SOM is able to tolerate missing data and some nodes might be empty due to incompletely defined data, it is desirable to have an average of at least 5–10 observations per node for statistical significance.[11] Hence, too many empty nodes may indicate that the map size is too big for the number of observations, and progressively reducing the grid size until only a handful of empty nodes are present would make it possible to reach a suitable map size. An alternative approach that works in the opposite direction is the growing self-organising map (GSOM), a growing variant of the SOM.[12] The GSOM was specifically developed to address the issue of identifying a suitable grid size in the SOM. It starts with a minimal number of nodes (usually four) and grows new nodes on the boundary based on a heuristic process. By selecting a value called the *spread factor* before training the model, the operator has the ability to control the growth of the GSOM. In addition, there are multiple ways of growing a SOM: either by readjusting the positions of the given neurons defined by the best matching unit (BMU) or by using newly produced neurons and assigning them to a suitable location. In sum, GSOM involves even more parameter tuning than normal SOM, offsetting the advantage of not needing to choose the map size. Moreover, only few implementations of the GSOM algorithm are currently available for R and Python; these implementations were evaluated by other authors[13] and the results were defined as unsatisfactory in terms of performance. A final approach consists of training the SOM at various grid sizes and evaluating each map's quality to ascertain the optimal map size.[14] Several measures can be computed to assess the quality of a SOM:

1. Quantization error (Qe), which measures map resolution
2. Topographic error (Te), which measures topology preservation (i.e., close observations in the original space should be mapped to close units in the SOM)
3. Kaski–Lagus error (KLe), which combines aspects of the quantization and topographic errors (for relatively small data sets, this metric is typically sufficient, obviating the need for additional measures[15])
4. Percentage of explained variance (%ev), which measures the proportion to which the model accounts for the variation of a given data set (higher values indicate better quality)

The characteristics of these measures are described in detail elsewhere.[7]

In this work, we propose a method that combines all these different approaches and identifies the optimal grid size automatically, so that the analyst does not need to specify the map size at the beginning of the process. First, the maximum size of a square grid (rectangular grids were not investigated in this study) is calculated using Equation (1). Then, a growing iterative process is executed, starting with a 2-by-2 hexagonal toroidal (i.e., the edges of the map are joined) grid and adding a new row and column at each step until the final map reaches the maximum size calculated beforehand. During this process, especially when working with particularly large data sets, the number of iterations of each training step are set to a relatively low value, for example, <500, to reduce computation time, while all the other arguments such as the learning rate are set to default. In addition, when computer performance is limited, it is possible to increment the grid size according to a specific sequence of numbers, for example, by two or five rows/columns each time. The four quality parameters are measured for each map produced during this process and each Qe, Te, KLe and %ev are normalised within their respective ranges of values. Next, a quality index (QI) that includes

all the above-mentioned normalised quality measures is calculated for each map using a heuristic formula as shown in Equation (2):

$$\mathrm{QI} = n\%\mathrm{ev} - (n\mathrm{Qe} + n\mathrm{Te} + n\mathrm{KLe}), \quad (2)$$

where n%ev, nQe, nTe and nKLe are the normalised values of the four quality measures. As a result, the map size that maximises the range between %ev and the sum of the three errors indicates the optimal compromise between all the quality measures of the SOM model, that is, the ideal grid size corresponds to the highest QI. Finally, the model is re-trained using exclusively the optimal grid size, this time with a greater number of iterations, for example, 500 or higher.

2.3.3 | Cluster analysis by HCA

Since SOM is primarily a data-driven dimensionality reduction and data compression method, and not a clustering technique, the nodes in the final map do not necessarily isolate groups of observations with similar metrics, especially when the number of SOM units is large. Therefore, clustering may be performed on the SOM nodes to group similar units and to facilitate quantitative analysis of the map and the data.[10] In general, clustering techniques such as *k*-means[16] present a problem that is similar to the manual identification of the optimal SOM size—a suitable number of clusters must be selected in advance and, in case, adjusted during a series of trials. Different approaches to tackling this problem are available. For example, an estimate of the number of clusters that would be suitable for discerning between similar node groupings can be determined using a *k*-means algorithm recursively for a range of cluster values and examining the plot of within cluster sum of squares (WCSS) for an elbow-point.[17] Alternatively, HCA, a method of cluster analysis that seeks to build a hierarchy of clusters,[18] can be used. The Kelley–Gardner–Sutcliffe (KGS) penalty function for a hierarchical cluster tree[19] allows the optimal number of clusters to be estimated automatically by selecting the lowest penalty. Ideally, the clusters identified are contiguous on the map surface, but this depends on the underlying distribution of variables.

2.3.4 | Cluster assignment

The two-stage procedure (first using SOM to produce the nodes that are then clustered by HCA in the second stage) described in the sections above was used to analyse the EDX and XRF data sets. In either case, after the clustering algorithm had been applied to the SOM map for assigning clusters to each of the nodes, the generated clusters were also assigned to the original observations in the data set, to produce plots in the form of time series or chemical maps.

3 | RESULTS AND DISCUSSION

3.1 | Preliminary data visualisation

In this study, SOM-HCA, a fully automated data mining method, was tested for the interpretation of large x-ray spectral data sets. By using the SOM method explained in section 2.3.2, maps of 3-by-3 nodes and 40-by-40 nodes were produced for the EDX and the XRF data sets, respectively. There are a number of different plot types available for visualising the quality of the generated SOMs and to explore the relationships between the variables in the data sets. The node count plot (Figure 3a,c) allows us to visualise how many observations are mapped to each node on the map, and can be used as a measure of map quality. Often referred to as the *U-Matrix*,[20] the neighbour distance plot (Figure 3b,d) shows the distance between each node and its neighbours—areas of low neighbour distance indicate groups of nodes that are similar, while areas with large distances indicate the nodes are much more dissimilar. Both metrics showed that the sample distributions in our data sets were relatively uniform, with no empty nodes, which would appear coloured in grey, and not too many large values. The EDX data set featured 22% of the nodes with more than 10 observations and 11% of the nodes with a neighbour distance greater than 400 units, while the XRF data set showed ~40% of the nodes with more than 100 observations and around 10% of the nodes with a neighbour distance >1000 units. The node weight vectors, or codes, are made up of normalised values of the original variables used to generate the SOM. Each node's weight vector is representative of the observations mapped to that node. By visualising the weight vectors across the map, patterns in the distribution of observations and variables can be observed. In the case of the EDX data set, these patterns corresponded to individual sum spectra that were associated with their respective nodes (Figure 4a). According to this visualisation, the signals of Ca, Zn and Pb appeared to be especially intense, and one node exhibited a notably high concentration of Ca. Finally, a SOM heat map (Figure 5) allows the visualisation of the distribution of a single variable across the map. Typically, a SOM investigative process involves the creation of multiple heat maps, and then the comparison of these heat maps to identify interesting areas on the map. It is worth noting

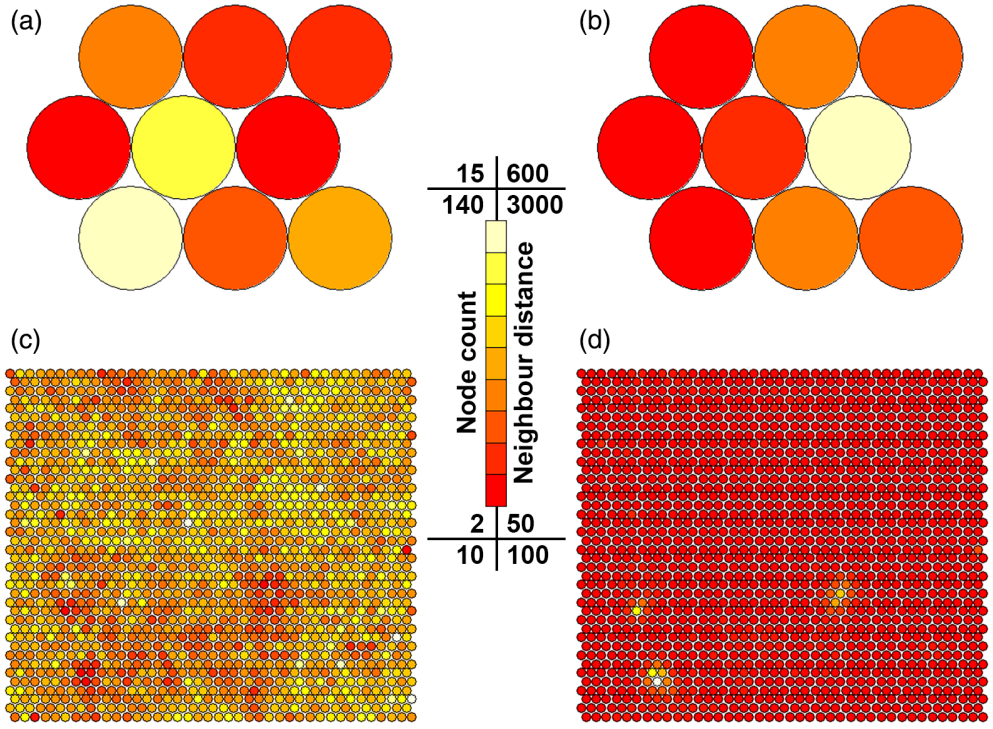


FIGURE 3 Node count plots (a and c) and neighbour distance plots (b and d) of the EDX data set (a and b) and XRF data set (c and d). The scale bar shows minimum and maximum values for each of the four plots. [Colour figure can be viewed at wileyonlinelibrary.com]

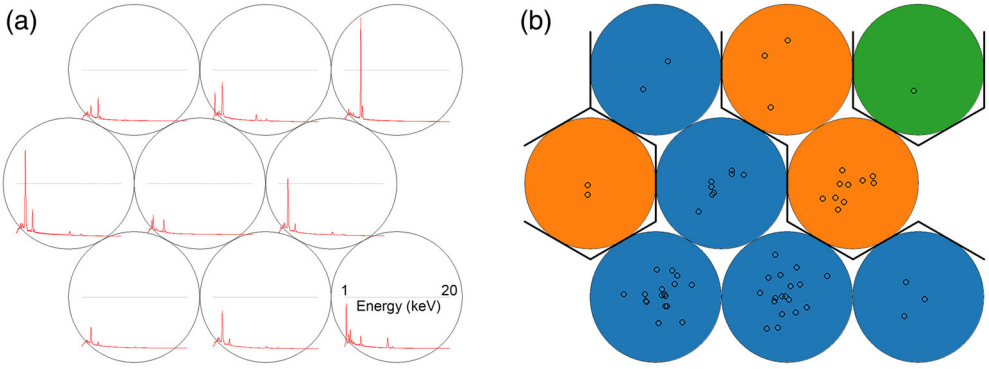


FIGURE 4 Node weight vectors (a) and assigned clusters (b) for the EDX data set. Each node in subfigure (a) displays a graphical representation of the magnitude of each variable (i.e., x-ray energy) in the weight vector; when the number of variables is high, the default visualisation of the weight vectors takes the form of a spectral pattern. The dots contained within each node in subfigure (b) represent the observations that have been mapped to that particular node. The equivalent plots for the XRF data set are not shown because of poor readability due to the higher number of nodes. [Colour figure can be viewed at wileyonlinelibrary.com]

that the individual observation positions do not change from one visualisation to another, the map is simply coloured by different variables. The heat maps generated from the XRF data set were compared with the corresponding elemental maps to validate a number of hypotheses. For example, the heat maps of P and Ca showed a direct relationship between those two elements, indicating the use of bone black (mostly $Ca_{10}(PO_4)_6(OH)_2$) in multiple areas of the painting; this assumption was confirmed by comparing the elemental distribution maps of P and Ca. Conversely, the heat map of Cu indicated that pigments containing this element were applied sparingly in specific areas, effectively assisting with the challenge of a poorly contrasted Cu distribution map. Moreover, although the elemental maps indicated a correlation between the signals of Co and Ni, the heat maps of these elements did not exhibit a similar relationship, showing that a Co-based pigment was applied in a very selective way and suggesting that the presence of Ni could potentially be an artefact. Finally, as evident from the elemental maps, it is often challenging to distinguish the signals of Ti and Ba. However, the heat maps revealed that the presence of Ti is significantly more prominent in specific, small areas of the painting.

The HCA algorithm was applied to the SOM models for assigning a specific cluster to each node, with the

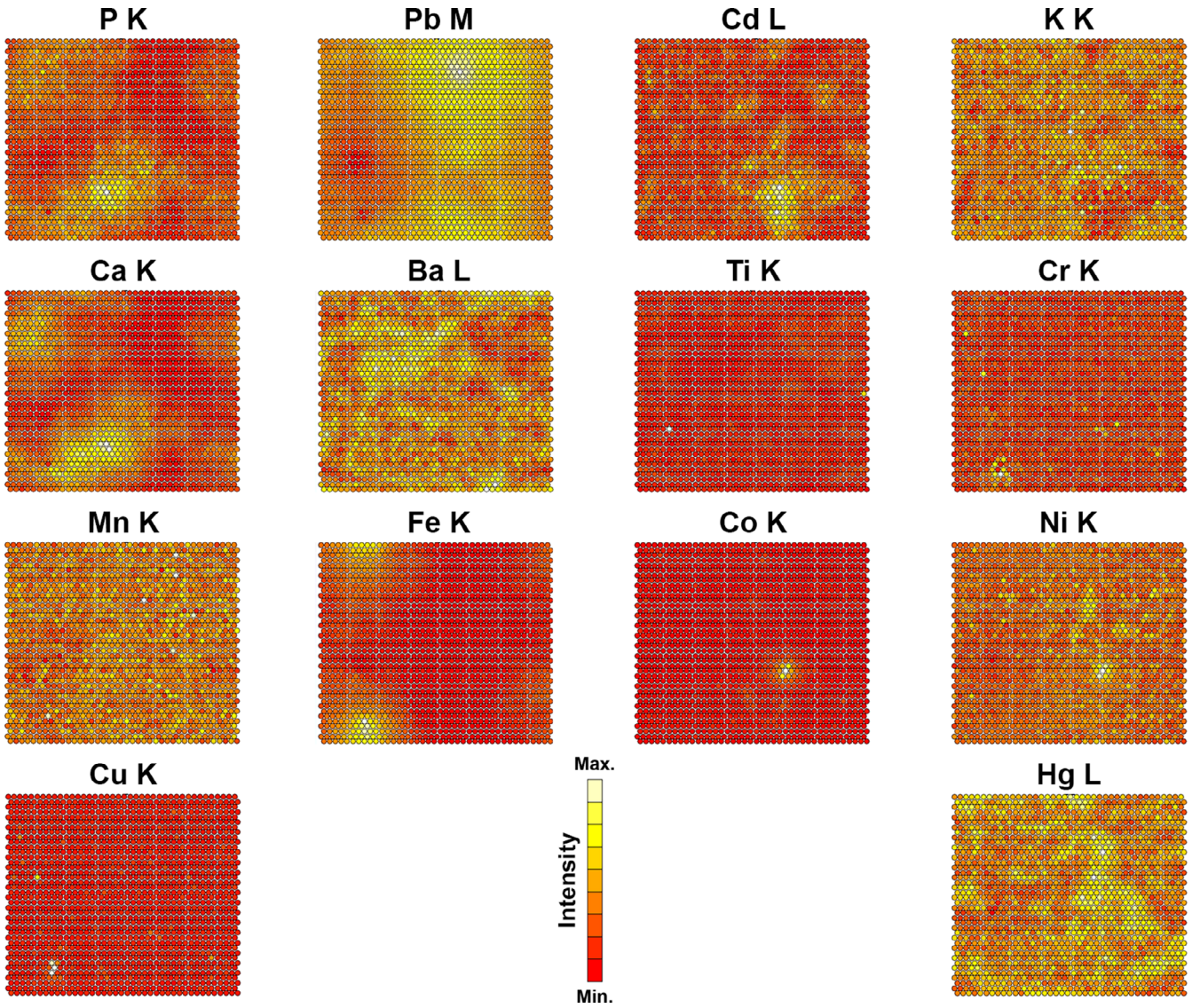


FIGURE 5 SOM heat maps showing the distribution of individual variables (i.e., elements spectral lines) across the XRF map. Heat maps related to the EDX data set are not shown due to the higher number of variables. [Colour figure can be viewed at wileyonlinelibrary.com]

number of clusters being determined based on the minimum value of the KGS function. In the case of the EDX data set, three clusters were used, while for the XRF data set, seven clusters were selected. These clusters were subsequently assigned to the original observations within each data set, allowing for exploratory visualisation. One noteworthy observation is that one of the three clusters in the EDX data set consisted of a single node (Figure 4b, green cluster), which corresponded to the weight vector showing a particularly strong Ca signal.

3.2 | EDX data visualisation

A sum spectrum was produced for each of the three EDX clusters, while the paintings grouped in each cluster were plotted against a timeline (Figure 6). In this way, the time distribution of the paintings containing specific types of ground materials could be visualised. The results showed that apart from cluster 3, in which the ground layer is composed exclusively of calcium carbonate, the other two clusters are characterised by lead white-based ground layers in combination with other materials such as zinc-based pigments (e.g., zinc white and lithopone), aluminosilicates and calcium carbonate. The difference between these two groups lies in the ratio between the spectral intensities of lead and the other elements, indicating that cluster 2 corresponds to a recipe with a relatively higher concentration of lead white compared with cluster 1. The majority of the paintings (44) aligned with cluster 1, while fewer works (17) correlated with cluster 2, suggesting Hammershøi's preference for the former type of ground

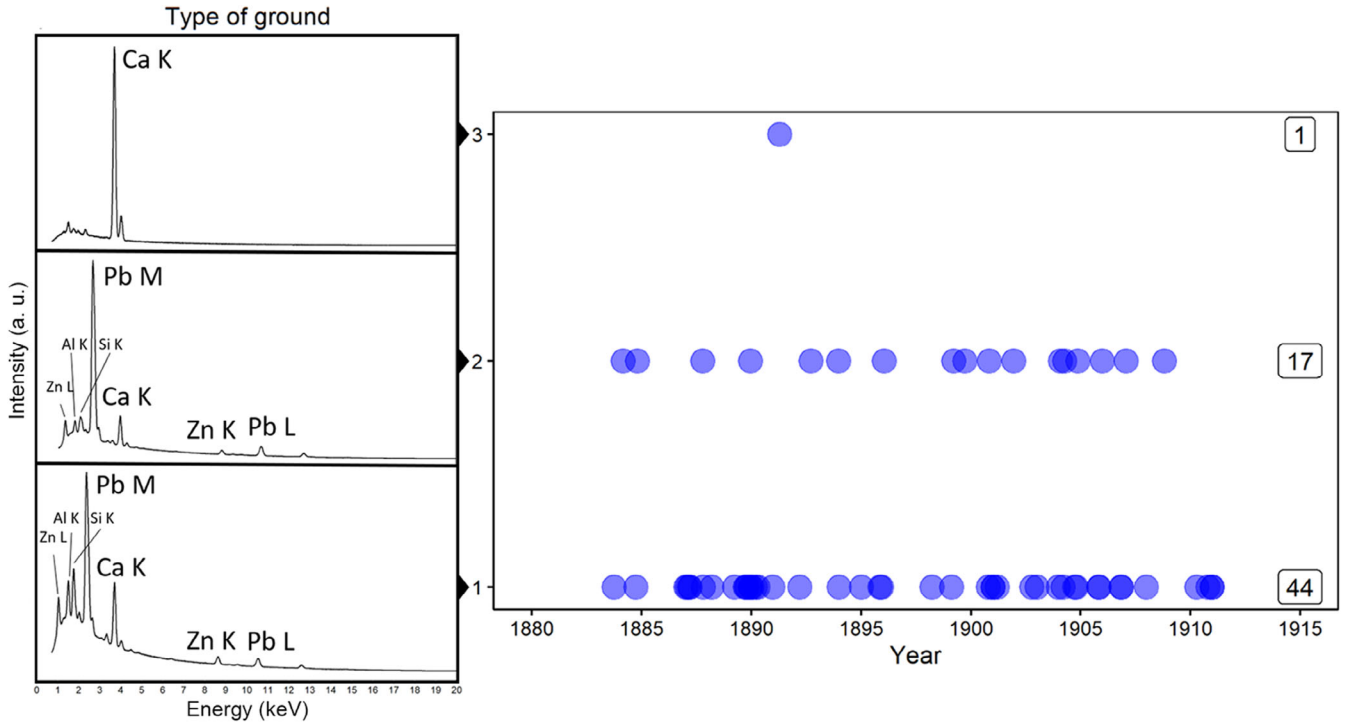
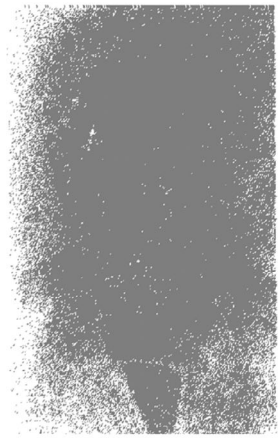
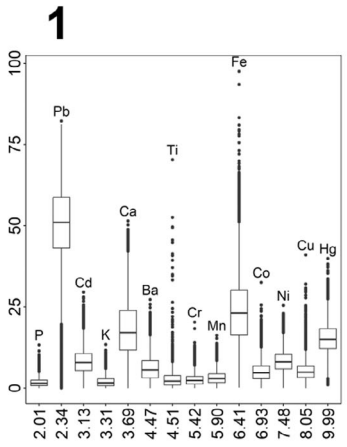


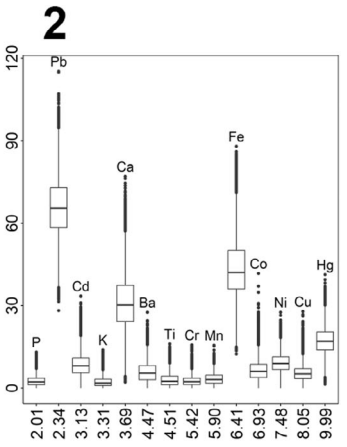
FIGURE 6 Sum spectra representing the composition of different ground types as determined by the calculated EDX clusters (left), alongside a chronological arrangement of paintings classified within each cluster (right). [Colour figure can be viewed at wileyonlinelibrary.com]

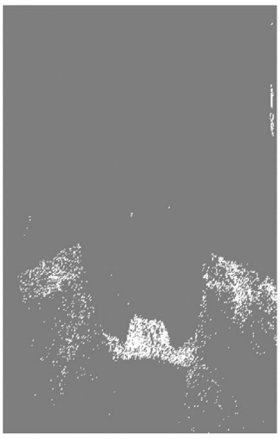
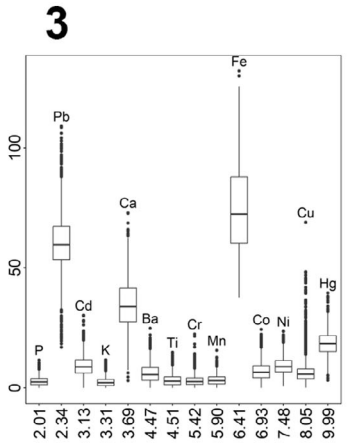
due to reasons that may range from availability to cost. However, there seems to be no clear boundary in terms of time between the use of one type of ground and the other, showing that both types were used concurrently throughout Hammershøi's career. Only one painting (a self-portrait executed in 1891) associated with cluster 3, differing from all other paintings of the set. This may be explained by noting that at that point of time, Hammershøi was based in Paris, where he may have had access to different materials. Without the proposed SOM-HCA method, reaching these conclusions would have required manually inspecting and comparing a large number of spectra, which would have been a time-consuming and laborious process. Our approach provided a more efficient and streamlined analysis, allowing us to focus on the most relevant information and resulting in a more accurate and reliable characterisation of the data.

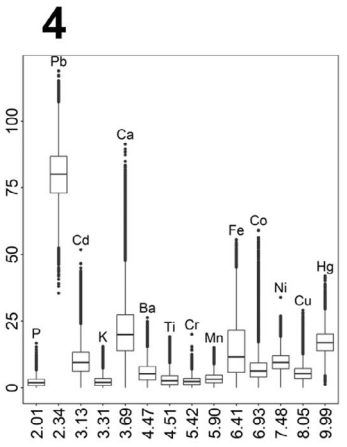
3.3 | XRF data visualisation

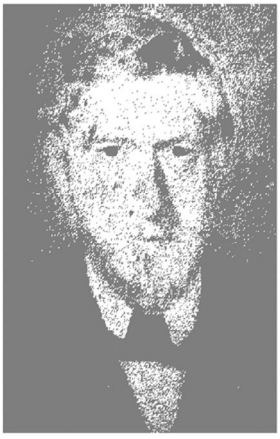
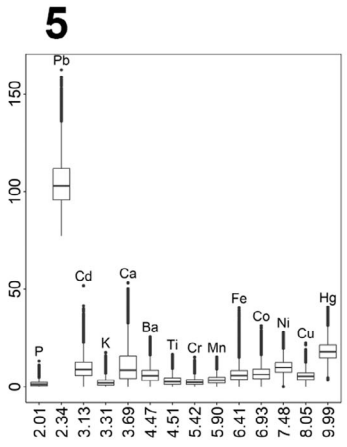
A different plotting method was used to automatically reduce the XRF data cube to a collection of distinct images (cluster maps) in which groups of pixels share similar spectra, making it possible to identify the materials that compose the paint layer more accurately. It is worth noting that the clustering output generated by the SOM-HCA method relies entirely on the statistical characteristics of the input data. Therefore, the resulting clusters may not always align with distinct partitions in the real-world context. One potential approach to overcome this limitation is to analyse the internal features and properties of the data elements that belong to each cluster. By conducting an in-depth examination of the intra-cluster characteristics, it is possible to identify underlying patterns or relationships that are not readily apparent from the clustering output alone. For that reason, each cluster map was accompanied by a box-and-whisker plot showing the relative abundances of the elements that characterise that particular cluster (Figure 7). The results indicated that the paint layer is composed of lead white ($2PbCO_3 \cdot Pb(OH)_2$, which is present in relatively high amounts in all clusters), bone black, iron-based earth pigments, vermilion (HgS), cadmium yellow (CdS) and cobalt blue ($CoO \cdot Al_2O_3$). Cluster 1 was mostly associated with parts of the background and the edges of the dark jacket, which were primarily produced with a mixture of earth pigments (containing Fe, but with very low levels of K and Mn, which may not be readily apparent from the elemental maps), bone black and vermilion. A similar composition characterised cluster 2, but the above-mentioned pigments were used in higher amounts to paint the darker areas of the jacket and the outline of the bowtie. The composition of the dark bowtie was

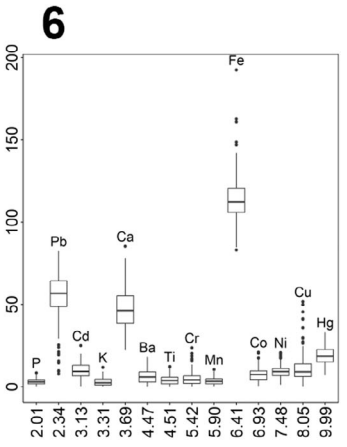



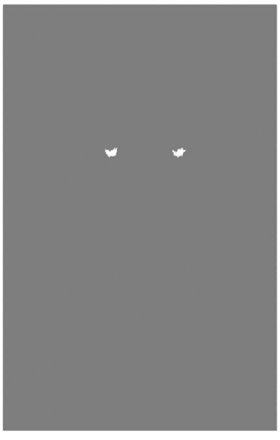
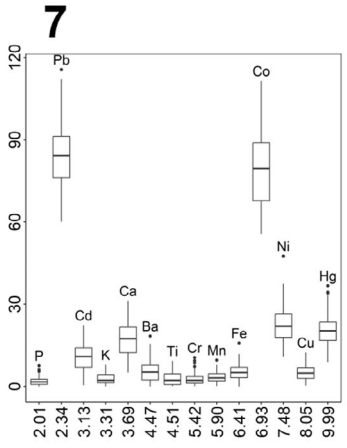



FIGURE 7 Legend on next page.

described by cluster 3, showing, apart from bone black, a substantial presence of Fe-based earth pigments. Cluster 4 showed that lead white mixed with variable amounts of bone black, cadmium yellow and cobalt blue is present in the hair and facial hair, in the shadow cast by the nose and in the background. Cluster 5 was dominated by Pb, which is present in great amounts as lead white in the off-white shirt and in the skin tone together with little amounts of vermilion. Cluster 6 was very similar in distribution and composition to cluster 3, but was correlated to retouching areas, as suggested by higher intensities of the signals of Fe, Cu and Cr. Cluster 7 was probably the most notable, since it only included the eyes of the figure, which were painted with abundant cobalt blue. Overall, the intensities of the different elements attributed to each cluster revealed that the muted hues in this composition, especially in the background and in the artist's outfit with their seemingly reduced tonality, were produced not just by mixing white, black and brown pigments, but also by adding other paints such as vermilion, cadmium yellow and cobalt blue. This is not very obvious in the single-element maps (Figure 2), where the broad dynamic range required for display of the distribution of an element results in the omission of low intensity areas. The cluster maps associated with the statistics of their respective chemical properties enabled us to identify the main elemental associations in a more efficient and accurate manner, without the need to manually overlay and compare a larger number of elemental maps. This provided a more comprehensive representation of the association of the different elements as well as a more precise characterisation of the data.

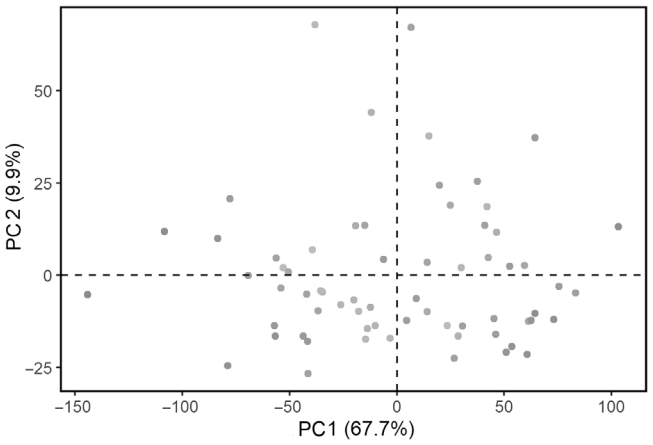


FIGURE 8 PCA score plot of the EDX data set. Percent values in parentheses represent explained variance.

3.4 | Comparison with other data mining techniques

One may wonder how these results compare to those obtained using other well-established data mining methods. Principal component analysis (PCA) and *k*-means, two methods for analysing large data sets containing a high number of features per observation, which are commonly used in heritage science studies[21,22] were tested on the EDX and XRF data sets, respectively. Unlike SOM-HCA, PCA did not increase the interpretability of the EDX data while preserving the maximum amount of information, as the visualisation of the multivariate data prevented any groups of specific ground materials from being identified (Figure 8). *K*-means cluster analysis generated a clusters map that is fairly similar to the SOM-HCA clusters map (Figure 9). However, as discussed above, the *k*-means algorithm requires the number of clusters to be specified in advance, which is considered one of the biggest drawbacks of this method; in this case, we used the number of clusters generated by the SOM-HCA method. Furthermore, the *k*-means method prefers clusters of approximately similar size, as it will always assign an object to the nearest centroid. This often leads to incorrectly cut borders of clusters, which in Figure 9 is very visible, for example, in the outlines of the facial features and the areas corresponding to the eyes.

4 | CONCLUSIONS

In this work, we aimed at developing and testing a fully automated data mining process for extracting and discovering patterns in large x-ray emission spectroscopy data sets obtained from SEM–EDX and MA-XRF analyses during a technical art historical systematic study of paintings by Vilhelm Hammershøi. The proposed method combines SOM, an artificial neural network-based type of unsupervised machine learning, and HCA, a cluster analysis technique, which were tweaked in such a way as to address the problem of specifying the number of nodes and clusters in advance. The method helped produce clusters related to specific chemical compositions, which

FIGURE 7 Cluster maps of the portrait of *Jens Ferdinand Willumsen* with corresponding elements spectral lines intensities. In each box-and-whisker plot, the *x*-axis represents energy (keV) and the *y*-axis represents net intensity (counts/1000 s). The box-and-whisker for each selected spectral line displays the median, interquartile range, minimum-maximum range, and outliers in its net intensity.

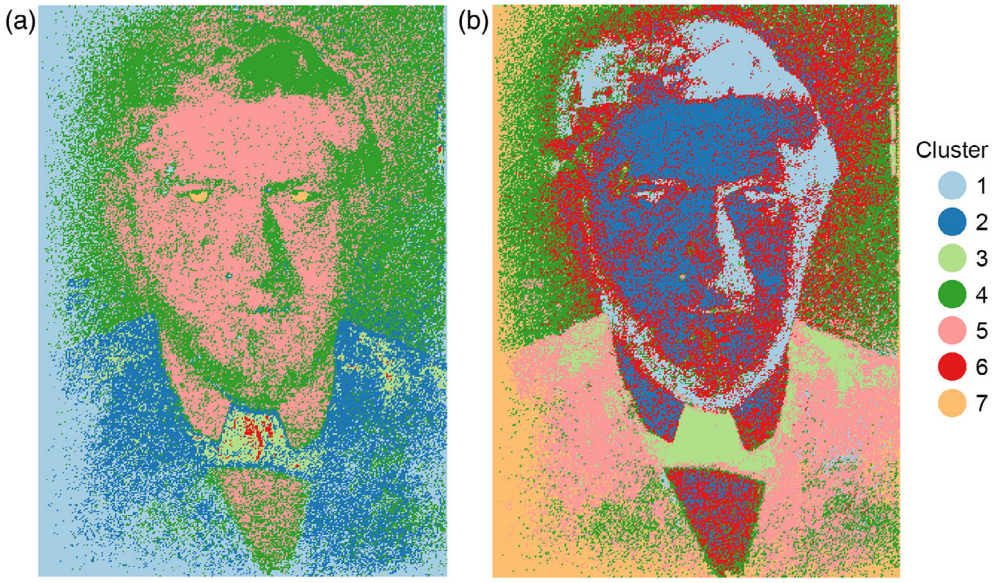


FIGURE 9 Comparison between the distributions of the clusters identified by SOM-HCA (a) and *k*-means (b) in the XRF data set. [Colour figure can be viewed at wileyonlinelibrary.com]

enabled efficient pigment identification in the ground layers of a large number of paintings as well as in the paint layer examined at the surface of one selected painting. The process was shown to perform well when compared with other well-established, but limited, data mining methods such as PCA and *k*-means, and it allowed reducing the time necessary for the interpretation of the results significantly. In summary, the proposed approach for interpreting both the EDX and XRF data sets showed considerable potential to enable automatic, accurate and time-efficient exploration of x-ray spectral data, which will facilitate the analysis of the extensive results collected during the ViHDA project, allowing for the generation of new information that will be made available in an open access digital resource. As SOM can also be used for supervised machine learning, future research will focus on validating the performance of the SOM-HCA method by creating simulated data sets based on mock-ups with well-defined properties. This will help in evaluating the accuracy and robustness of the technique further, and may lead to the development of even more efficient and reliable versions of the proposed method. The next phases of the ViHDA project will also investigate the use of the proposed SOM-HCA method on other types of data, such as complementary spectral data in the ultraviolet–visible-infrared range for the analysis of organic components, to provide more comprehensive details regarding Hammershøi's technique across all the examined artworks. In addition to finding trends in the use of materials, it will be possible to correlate the obtained chemical information with other properties observed on Hammershøi's paintings such as formats, canvas thread counting, state of preservation and composition. Since such evidence-based knowledge will be essential not only to understand the significance of Hammershøi's works, but also to determine authenticity, provenance and dating, this information will be a valuable and necessary resource for assisting museums and private individuals in the assessment of the quality and geographical/temporal location of artworks that could have originated from Hammershøi's hand.

ACKNOWLEDGEMENTS

The authors are grateful to Troels Filtenborg, Sofie Wikkelsø Jensen and Oscar Holm (National Gallery of Denmark) for technical assistance in the collection and preparation of the samples used in this research, and the Royal Danish Academy – Institute of Conservation for the use of the SEM–EDX instrumentation. The authors also thank Pauline Lehmann Banke and Loa Ludvigsen (National Gallery of Denmark) alongside independent researchers Jørgen Wadum and Annette Rosenvold Hvidt for discussions. Finally, the authors gratefully acknowledge funding from the Augustinus Foundation and the New Carlsberg Foundation.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Gianluca Pastorelli https://orcid.org/0000-0001-6926-1952
Annette S. Ortiz Miranda https://orcid.org/0000-0001-9223-8099
Anne Haack Christensen https://orcid.org/0000-0001-7934-8832

REFERENCES

[1] P. Vad, K. Tindall, *Vilhelm Hammershøi and the Danish Art at the Turn of the Century*, Yale University Press, New Haven **1992**.

[2] A. Bintintan, M. Gligor, I. D. Dulama, S. Teodorescu, R. M. Stirbescu, C. Radulescu, *Rev. Chim.* **2017**, *68*, 847.

[3] V. Renda, V. M. Nardo, G. Anastasio, E. Caponetti, C. S. Vasi, M. L. Saladino, F. Armetta, S. Trusso, R. C. Ponterio, *Spectrochim. Acta, Part B* **2019**, *159*, 105655.

[4] S. Kogou, L. Lee, G. Shahtahmassebi, H. Liang, *X-Ray Spectrom.* **2021**, *50*, 310.

[5] R. Wehrens, L. M. C. Buydens, *J. Stat. Softw.* **2007**, *21*(5), 1. https://doi.org/10.18637/jss.v021.i05

[6] R. Wehrens, J. Kruisselbrink, *J. Stat. Softw.* **2018**, *87*(7), 1. https://doi.org/10.18637/jss.v087.i07

[7] J. Boelaert, E. Ollion, J. Sodoge. aweSOM: Interactive Self-Organizing Maps [Online]. https://CRAN.R-project.org/package=aweSOM (accessed: 16 June 2023).

[8] D. White, R. B. Gramacy. maptree: Mapping, Pruning, and Graphing Tree Models. 2022 [Online]. https://CRAN.R-project.org/package=maptree (accessed: 16 June 2023).

[9] T. Kohonen, *Self-Organizing Maps*, Vol. 30, Springer Science & Business Media, Heidelberg **2012**.

[10] J. Vesanto, E. Alhoniemi, *IEEE Trans. Neural Netw.* **2000**, *11*, 586.

[11] T. Kohonen, *MATLAB Implementations and Applications of the Self-Organizing Map*, Vol. 177, Unigrafia Oy, Helsinki, Finland **2014**.

[12] S. Delgado, C. Gonzalo, E. Martínez, A. Arquero. in *IGARSS 2004. 2004 IEEE Int. Geosci. Remote Sens. Symp.*, Vol. 1, 2004.

[13] R. S. Adeu, K. R. Ferreira, P. R. Andrade, L. Santos. in *Proc. XX GEOINFO*, November 11–13, 2019, SP, Brazil 2009.

[14] D. L. B. Fortela, M. Crawford, A. DeLattre, S. Kowalski, M. Lissard, A. Fremin, W. Sharp, E. Revellame, R. Hernandez, M. Zappi, *Clean Technol.* **2020**, *2*, 156.

[15] S. Kaski, K. Lagus, *ICANN* **1996**, *96*, 809.

[16] E. W. Forgy, *Biometrics* **1965**, *21*, 768.

[17] A. Alghamdi, G. Hu, H. Haider, K. Hewage, R. Sadiq, *Sustainability* **2020**, *12*, 4422.

[18] F. Nielsen, *Introduction to HPC with MPI for Data Science*, Springer, Cham **2016**, p. 195. https://doi.org/10.1007/978-3-319-21903-5

[19] L. A. Kelley, S. P. Gardner, M. J. Sutcliffe, *Protein Eng., des. Sel.* **1996**, *9*, 1063.

[20] A. Ultsch, in *Proc. Int. Neural Netw. Conf. (INNC-90), Paris, France*, Vol. 1 (Eds: B. Widrow, B. Angeniol), Kluwer, Dordrecht, Netherlands **1990**, p. 305.

[21] M. Albrecht, O. de Noord, S. Meloni, A. van Loon, R. Haswell, *Herit. Sci.* **2019**, *7*(1), 1.

[22] H. Chopp, A. McGeachy, M. Alfeld, O. Cossairt, M. Walton, A. Katsaggelos. Denoising fast x-ray fluorescence raster scans of paintings [Online]. http://arxiv.org/abs/2206.01740 (accessed: 16 June 2023).

How to cite this article: G. Pastorelli, A. S. Ortiz Miranda, A. H. Christensen, *X-Ray Spectrom* **2024**, *53*(5), 392. https://doi.org/10.1002/xrs.3388