IOP Conf. Series: Materials Science and Engineering **364** (2018) 012045 doi:10.1088/1757-899X/364/1/012045

A data mining approach for indoor air assessment, an alternative tool for cultural heritage conservation

D Leyva Pernia[1,4,5], **S Demeyer**[1], **O Schalm**[2], **W Anaf** [3]

[1] Department of Mathematics - Computer Science, University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium
[2] Conservation Studies, University of Antwerp, Blindestraat 9, 2000 Antwerp, Belgium
[3] War Heritage Institute, Renaissancelaan 30, B-1000 Brussels, Belgium
[4] Department of Physics, CEADEN, 502, Calle 30, 11300 Havana, Cuba

[5] Diana.LeyvaPernia@uantwerpen.be

Abstract. The exposure of cultural heritage to the environment has a significant impact on its degradation process and degradation rate. Consequently, managing the indoor air quality is vital to minimize further damage to historical artefacts and works of art. Despite its potential impact, the traditional assessment of the indoor air quality still represents a challenge for most collection guardians. This approach typically relays on the comparison of measured environmental parameters and corresponding acceptable values. However, determining the acceptable values and relative importance of the different environmental parameters turns out to be quite complex since it depends on the material types present in the collection and their preservation state. Furthermore, the significant amount of data generated during the measurements hampers the application of traditional methods of analysis.
Considering all these, we propose the use of data mining as an alternative method for the indoor air quality assessment in cultural heritage studies. Data mining can provide knowledge from vast volumes of heterogeneous data, through high-speed processing, detection, and analysis. Here we present its application to identify dynamics and patterns affecting the indoor air quality in a realistic case. Using data from a measuring campaign held at a late Gothic church in Belgium, we show that inappropriate periods can be identified without using standards. In addition, different types of periods can be identified by studying the relation between multiple parameters. For that we use the k-means clustering method, interpreting the results with both visual and statistical tools.

1. Introduction

Several studies have already demonstrated how the exposure to an aggressive environment has a significant impact on the degradation process and degradation rate of cultural heritage objects [1]. Based on these findings heritage guardians and institutions have now an increasing interest in managing the indoor air quality and provide stable environmental conditions to minimize, as much as possible, the occurrence of further damage to historical or valuable artefacts on their collections. However, monitoring and evaluating relevant environmental data is a challenging task for most heritage guardians, mainly due to the considerable amount of information generated, and the uncertainties behind the selection of the acceptable environmental values.

Continuous indoor air quality monitoring implies the collection of vast amounts of data that can quickly become overwhelming for specialists with limited experience in data science. This is aggravated when an increasing number of environmental parameters are analysed. Temperature and relative

humidity are the most common parameters monitored by the heritage community, but other parameters such as visible light, UV radiation or pollution also have a remarkable influence on the degradation rate of most objects. With the currently available technology these parameters can be monitored as well, and they should be considered to achieve an accurate assessment [2].

Traditional assessments usually rely on the comparison of measured environmental parameters with their corresponding acceptable values. Unfortunately, the target values are not precisely known, and they depend on variables such as the material type, preservation state, etc. This entails a strong dependency on the selected guidelines that define acceptable environmental values, and will be reflected in the accuracy of the analysis [3].

Given the preceding, we propose a complementary approach to aid indoor air assessment for heritage conservation based on the implementation of Big-Data related techniques, such as data mining and data analytics. The goal of these techniques is to extract relevant knowledge from data. In our specific context we focus on interesting patterns or atypical behaviours [4]. A similar approach has already been successfully implemented in indoor environment quality assessment for human health and comfort [5]. However, the application of this method for a comprehensive environmental characterization is still uncommon in the field of cultural heritage studies.

We present here our results from the implementation of the techniques mentioned above for recognizing inappropriate periods in the absence of standards or guidelines, and identifying dynamics or patterns affecting the indoor air quality in a specific case study.

2. Materials and Methods

2.1. Case study

To evaluate the viability of our approach, we analysed the data gathered from 23/07/2017 to 23/10/2017 in a measuring campaign held at a late Gothic church in the centre of a small Belgian city. The environmental information was collected with an innovative multi-sensor tool that registered every 15 minutes the values of temperature, relative humidity, illuminance, particulate matter (PM), nitrogen dioxide (NO_2) and ozone (O_3) concentrations among other parameters [2]. This resulted in a data matrix of 8928 measuring points (matrix rows) and 7 measurements per point including the above mentioned parameters (matrix columns). In addition, every measuring point was labelled with a timestamp and the corresponding weekday number in which the measurements were performed (e.g., Monday = 1).

2.2. Identifying periods of elevated risk

Our method for the identification of periods of elevated risk is based on three main suppositions.

1. A stable climate is more conducive for heritage conservation, especially in the case of temperature and relative humidity acting over hygroscopic materials [6].
2. Keeping the light exposure and the concentration of pollutants as low as reasonable achievable minimises the risk of material degradation [7].
3. Short-time fluctuations will not necessarily affect all objects, since the reaction time of the piece must be shorter that the fluctuation in order to have a potential impact on its conservation state. [8]

Considering the third supposition, we start the analysis by filtering out the short-time fluctuations. There are several classes of filtering methods, but the most widely used apply statistical measures over moving windows [9]. The mean value is often used as the statistical measure for the data in the window. Consequently, a moving mean is calculated creating series of averages of different subsets as the window shifts over the full data set.

Moving means are ordinarily used to smooth out short-term fluctuations and highlight longer-term trends or cycles in time series data. In our case, the use of the moving mean method filters out the fluctuations caused by the day and night variations exposing the regular seasonal variations or the anomalous behaviour that occurred during periods of elevated risk.

Considering now the first and second suppositions we can establish that the periods where the filtered data presented the higher fluctuations were the periods most likely to be linked to an elevated risk of degradation. For temperature and relative humidity, the relevant fluctuations can be found above or below the mean value, while for the rest of the environmental parameters, only the fluctuations above the mean would be noteworthy.

2.3. Clustering data, k-means clustering

Clustering is an unsupervised learning method that assigns labels to objects in unlabelled data that can be implemented by different types of algorithms [9]. For our research, we selected the k-means clustering, a data-partitioning algorithm based on iterative minimization of the sum-of-squares criterion [10].

Clustering validation was performed with two different methods: the Davies-Bouldin index, and Silhouette value. The Davies-Bouldin criterion is essentially a ratio of within-cluster and between-cluster distances. Based on its analysis the optimal clustering solution has the smallest Davies-Bouldin index value [11]. The Silhouette value is a measure of how similar a point is to points in its own cluster when compared to points in other clusters [12]. This value ranges from -1 to +1, where a high silhouette value indicates that the point is well-matched to its cluster and poorly-matched to neighbouring clusters.

3. Results and Discussion

The complete data set studied is presented in Figure 1. As mentioned in the previous section, seven different environmental parameters were monitored every 15 minutes during 93 days, resulting in data streams of 8928 elements each. This data set cannot be regarded as considerably large when compared to a typical climate characterization campaign that should cover, at least, one full year of data. However, it is more than enough to hinder a prelaminar assessment through visual inspection. There are certainly patterns in the data, but there is too much information to get a clear grasp of them. One characteristic that all the parameters share is the unremitting presence of short time fluctuations, mainly related to the day-night cycle.

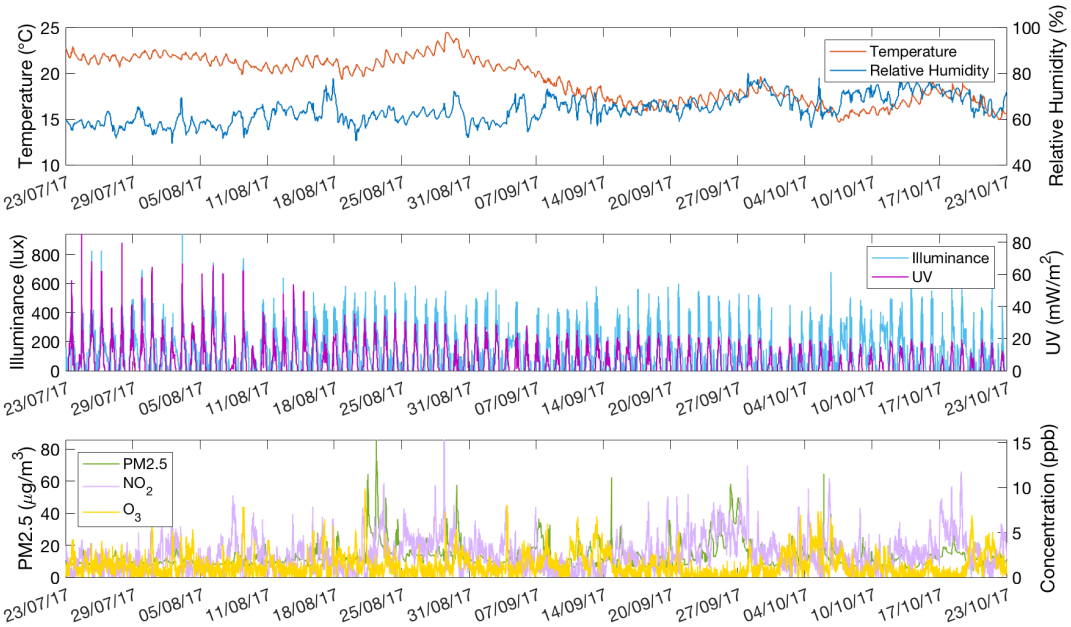

Figure 1. Environmental data registered for temperature, relative humidity, illuminance, ultraviolet radiation, particulate matter ($PM_{2.5}$), NO_2 and O_3 concentration.

In order to filter out the daily fluctuations we calculated the moving mean of the parameters with a symmetric window of 96 data points, which covers a period of 24 hours in the time scale of our case study. The result of its application on the case of relative humidity is presented in figure 2 (a). In this figure we plotted the measured data points for relative humidity, the mean value over the complete period analysed, the confidence bars delimited by the standard deviation (std) and the correspondent moving mean. By filtering higher frequency fluctuations, the moving mean shows now a gradual, but still appreciable overall increment during the studied period, which match the expectations for the seasonal transition from summer to autumn.

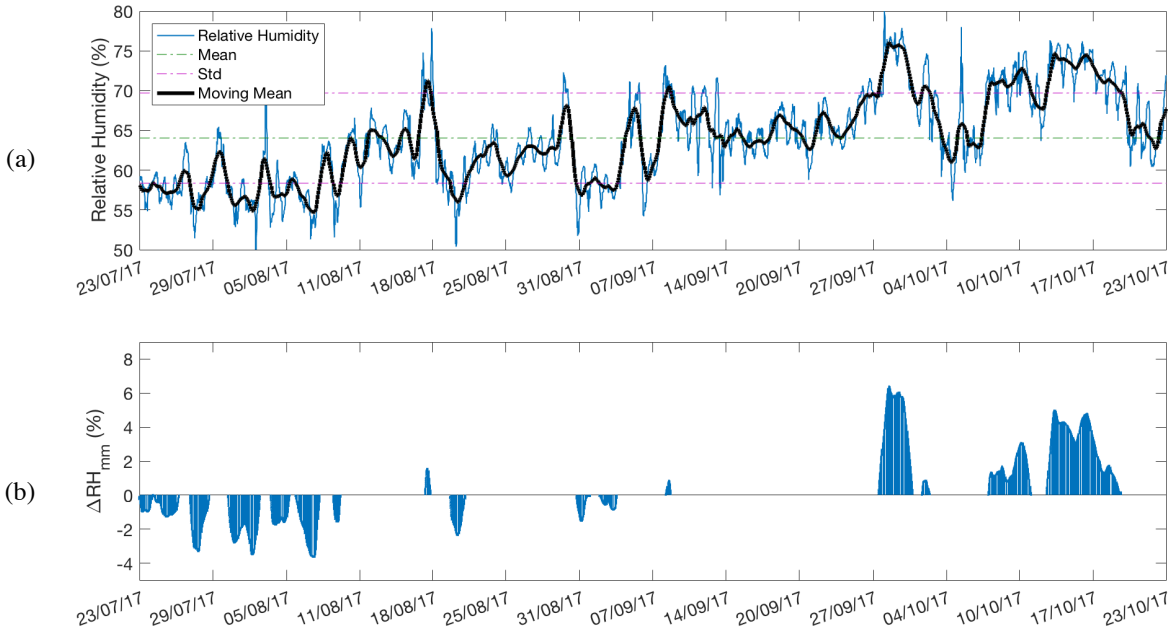

Figure 2. (a) Data registered for relative humidity, its mean value over the time analysed, confidence bar of ± the standard deviation, and moving mean for a symmetrical window of 24 hours. (b) Difference between the values of the moving mean outside of the confidence bar.

In figure 2(a) we can notice periods where the moving mean reached extreme values crossing the confidence bar. These periods are the most deviated from the typical behaviour, and consequently, more likely to be linked to unusual events or even to potentially dangerous occurrences. The magnitude of the difference between these points and the standard deviation can be used as a metric to characterize their anomaly. Plotting this magnitude would filter out the periods where the environmental parameter exhibits a standard behaviour, and therefore, directly guiding the analysis to periods of elevated risk. This analysis is illustrated in figure 2(b), where the difference between the values of the moving mean outside of the confidence bar and the standard deviation (ΔRH_{mm}) are plotted as a function of time.

This analysis facilitates the detection of extreme periods and can be implemented for all the environmental parameters on our dataset. Figure 3 shows the normalised values of these differences between the moving means and the corresponding confidence bars (y-axis) as a function of time for all parameters. The normalization was performed to establish an unbiased comparison unrelated to the measuring units of the parameters. For each case, the higher the difference value, the more atypical is the corresponding period. Due to the significant amount of data filtered out, it is still possible to identify the atypical episodes clearly, even when showing the information of all the parameters in the same graph.

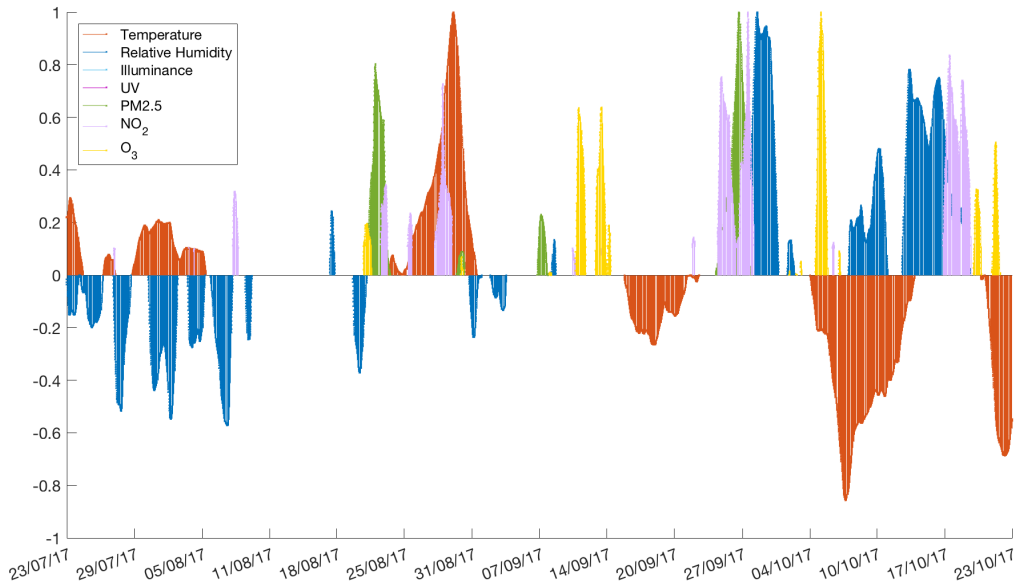

Figure 3. Normalised differences between the moving means of each environmental parameter and the corresponding confidence bars, identifying atypical periods.

For the implementation of the k-means clustering method, we analysed the entire data matrix of environmental data collected. Each column was normalised by rescaling in the range from 0 to 1 before applying the clustering method to avoid assigning artificial weights to the data. The clustering algorithm was implemented for k values from 2 to 15, each done in 5 separated runs. The resulting 13 cluster distributions were evaluated using the Davies-Bouldin and Silhouette criteria, and a detailed inspection of their characteristics. Based on these, we selected k=4 as the most relevant distribution for classifying our data set.

Table 1. Cluster interpretation data.

		Day	Temperature (ºC)	RH (%)	Illuminance (lux)	UV (mW/m^2)	Pm2.5 ($\mu g/m^3$)	NO_2 (ppb)	O_3 (ppb)
Cluster 1	min	1	19.153	51.436	0.000	0.000	5.153	0.000	0.000
	max	7	24.282	74.018	357.674	22.470	64.358	15.323	7.999
	mean	**3.862**	**21.244**	**61.026**	**40.040**	**1.935**	**12.406**	**2.754**	**1.165**
Cluster 2	min	1	14.646	56.147	0.000	0.000	2.973	0.000	0.000
	max	7	21.237	79.006	680.531	25.351	64.605	11.779	7.358
	mean	**4.045**	**17.205**	**65.755**	**236.407**	**8.743**	**14.333**	**2.355**	**1.613**
Cluster 3	min	1	14.647	60.476	0.000	0.000	4.024	0.000	0.000
	max	7	20.950	80.067	279.758	18.022	62.399	12.423	7.389
	mean	**4.063**	**17.008**	**69.350**	**30.157**	**1.148**	**11.937**	**3.086**	**1.146**
Cluster 4	min	1	18.163	49.346	0.000	0.000	6.164	0.000	0.000
	max	7	24.462	74.655	942.825	85.150	85.777	10.240	9.895
	mean	**4.088**	**21.646**	**59.063**	**284.142**	**17.988**	**12.891**	**1.461**	**1.614**

The data interpretation of each cluster is presented in Table 1. We found two sets of patterns that provide physical meaning to the cluster distribution. One pattern is defined by the presence of lower values of illuminance and ultraviolet radiation (clusters 1 and 3) or higher values of these two

environmental parameters (clusters 2 and 4). The other pattern is set by the anti-correlation of temperature and relative humidity. Clusters 1 and 4 hold higher temperature and lower relative humidity (summer), while clusters 2 and 3 are the opposite (autumn).

Plotting the temporal distribution of the 4 clusters, figure 4(a), confirms our hypothesis on the seasonal influence. Comparing the graphs in figure 4(a) and 4(b) we can confirm that clusters 1 and 4 comprehend the periods where the temperature was above the mean value, while the temperatures in clusters 2 and 3 where below the average. From figure 4(a) we can also conclude that the distribution based on the illuminance and UV levels does not exactly reflect the day-night variations. The number of days during the analysed period does not match the number of repetitions of the clusters, therefore the difference comes from the distribution of brighter days (clusters 2 and 4) and nights or darker days (clusters 1 and 3).

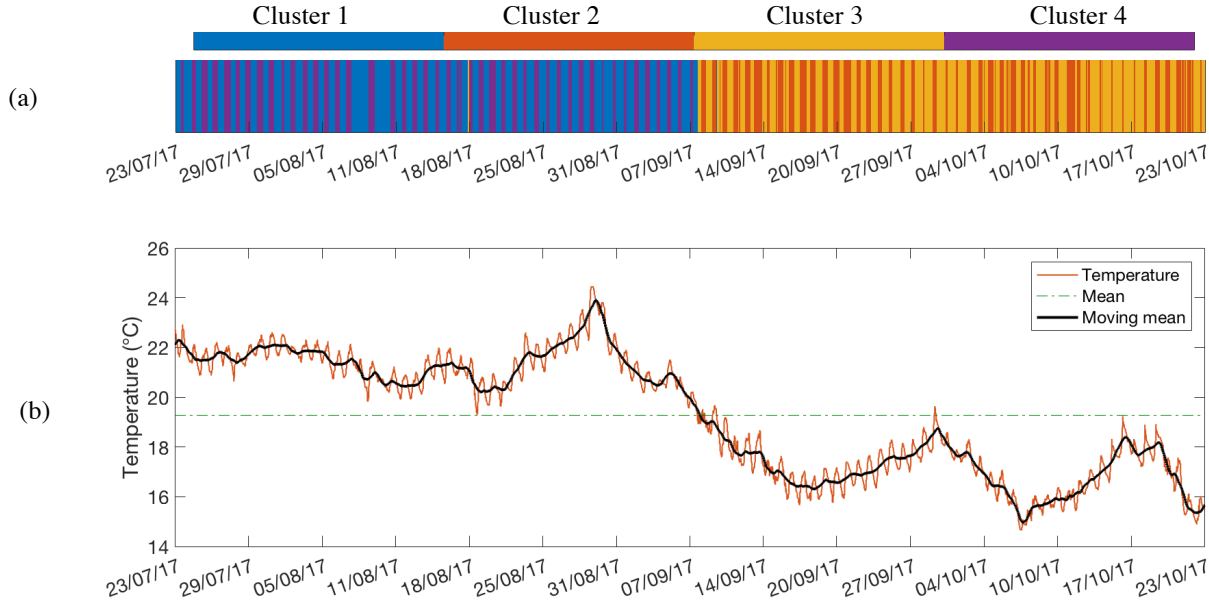

Figure 4. (a) Cluster distribution over time and (b) temperature distribution for comparison.

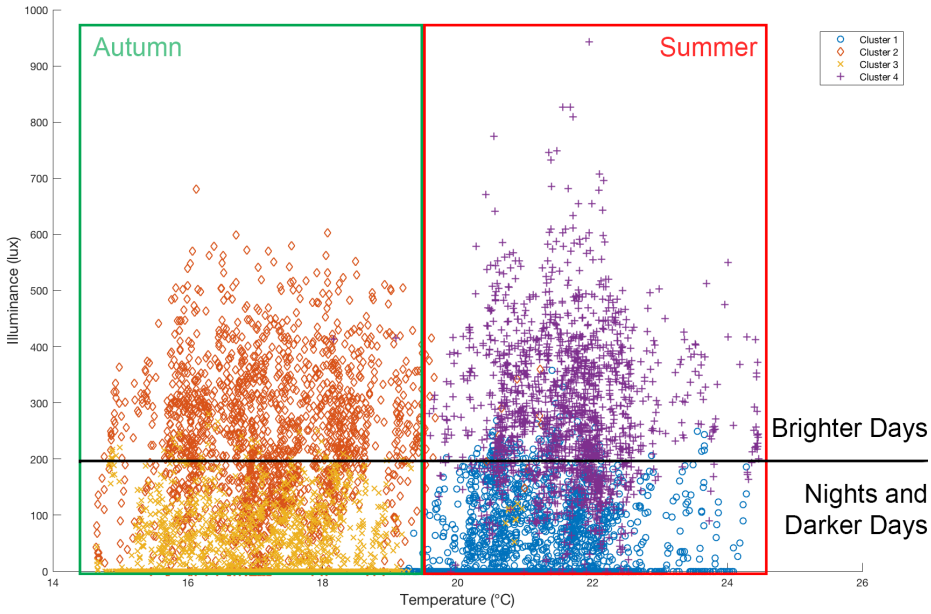

Figure 5. Cluster distribution for the plane temperature vs illuminance.

These distributions are also recognisable by plotting the clusters in the plane of temperature and illuminance (figure 5). In this graph the clusters can be found in 4 different quadrants representing nights and darker days of summer (cluster 1), brighter days of summer (cluster 4), brighter days of autumn (cluster 2), and nights and darker days of the same season (cluster 3); confirming our previous interpretation.

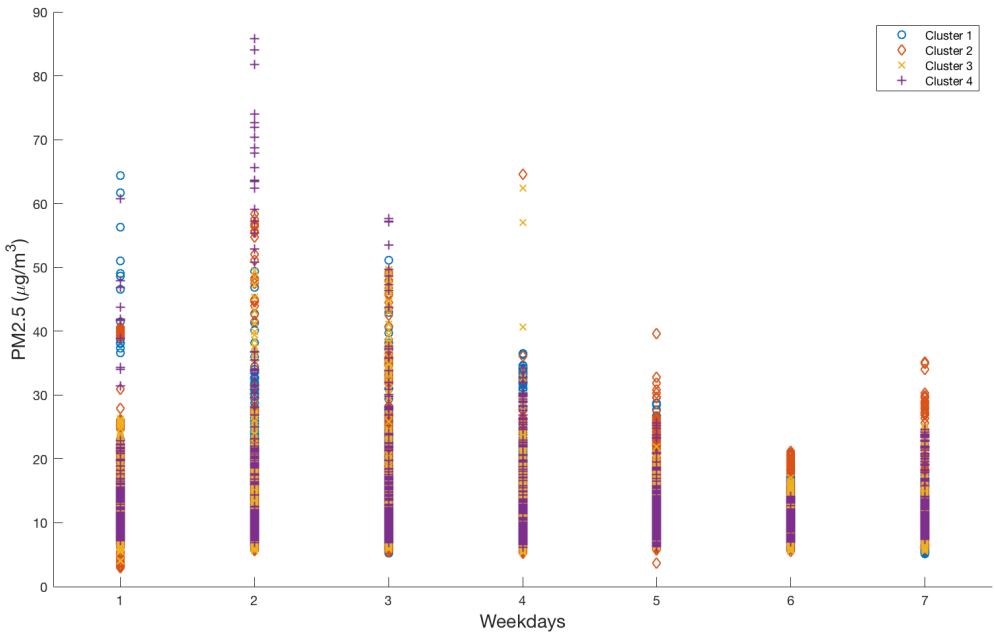

Figure 6. Cluster distribution regarding particulate matter and corresponding weekdays.

The analysis of the cluster distributions over different axes combinations can provide information about patterns in the environmental conditions. An example of this is presented in figure 6, where we plot the behaviour of particulate matter for each cluster as a function of the weekday number. In this figure is observed that, independently of the season, the lowest values of particulate matter were reached during the weekends, especially on Saturdays. In contrast, the higher concentrations, linked to periods of elevated risk, were reached on Tuesdays during brighter summer days (cluster 4). The source of this increment in the particulate matter concentration is most likely related to the construction works going on during the period. However, we cannot discard the influence of other sources, such as heavier traffic on workdays compared to weekends; or different internal sources, like a larger number of visitants during summer holidays. Nevertheless, being able to identify that this particular behaviour is connected to specific weather conditions can facilitate the selection of mitigation actions and alert heritage guardians about the potential risk during periods with similar characteristics.

4. Conclusions

This study revealed that, on the basis of three simple suppositions, it is possible to identify periods of elevated risk without the use of any specific guideline. The proposed approach detects periods of sudden change that deviates from the standard behaviour and filters out the information regarding stable periods. Despite this, the method would fail for periods of constant inappropriate environmental conditions. For that reason, we consider it as a complementary method, and not a substitution, to the use of norms or guidelines. However, in the cases where the application of direct guidelines is not a viable option, heritage guardians could still use our method to propose mitigation actions that would reduce the risk of degradation during the most extreme periods.

Our study also showed that the extended data matrix of environmental measurements does contain different types of patterns superposed on top of each other. By using clustering techniques, such patterns become easier to detect and interpret, facilitating the study of potential hazards.

Acknowledgements
The authors thank the financial support of the Belgian Federal Public Planning Service Science Policy (BELSPO) under project number BR/132/A6/AIRCHECQ.

References

[1] Thomson G 1986 *The Museum Environment*. Second edition ed.; Elsevier, Butterworth Heinemann: Oxford.

Caple C 2011 *Preventive Conservation in Museums*. Routledge, Taylor&Francis Group: London.

[2] Schalm O, Anaf W, Callier J and Leyva Pernia D 2018 New generation monitoring devices for heritage guardians to detect multiple events and hazards. *Proc. Int. Conf. The Future of Heritage Science and Technologies*, (Florence: Florence International Fair for Art and Restoration).

[3] Massimo A, Coppola F and Pavlovic A 2015 Application of the quality norms to the monitoring and the preventive conservation analysis of the cultural heritage. *International Journal for Quality Research*, 9: 299-308.

Leyva Pernia D, Demeyer S, Schalm O, Anaf W and Meert C 2016 New approach to indoors air quality assessment for cultural heritage conservation, *Proc. 14th Int. Conf. on Indoor Air Quality and Climate*, (Ghent: International Society of Indoor Air Quality and Climate); pp 490-497.

[4] Fayyad U M, Piatetsky-Shapiro G, Smyth P and Uthurusamy R 1996 *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park.

[5] Saarikoski M 2016 *A data mining approach to indoor environment quality assessment: A study on five detached houses in Finland*. MSc Thesis. Environmental Science. University of Eastern Finland; Department of Environmental and Biological Sciences.

[6] EN 15757:2010 *Conservation of Cultural Property – Specifications for Temperature and Relative Humidity to Limit Climate-Induced Mechanical Damage in Organic Hygroscopic Materials*, European Committee for Standardization, Brussels.

[7] Michalski S 1997 *The lighting decision*. In Textile Symposium 97, 97-104. Ottawa, Canada: Government of Canada.

ASHRAE 2011 Museums, Galleries, Archives, and Libraries. In *ASHRAE Handbook - Heating, Ventilating, and Air-Conditioning Applications (I-P Edition)*, American Society of Heating, Refrigerating and Air-Conditioning Engineers.

[8] Michalski S 1993 *Relative Humidity: A Discussion of Correct/Incorrect Values*. In ICOM Committee for Conservation.

[9] Runkler T A 2016 *Data Analytics*. Springer, Wiesbaden.

[10] Davies D L and Bouldin D W 1979 A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. PAMI-1, No. 2, pp. 224–227.

[11] Lloyd S 1982 Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*. Vol. 28, pp. 129–137.

[12] Rouseeuw P J 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. Vol. 20, No. 1, pp. 53–65.