ICOM-CC
18th Triennial Conference
2017 Copenhagen

PREVENTIVE CONSERVATION

Data mining in collections: From epidemiology to demography

CRISTINA DURAN*
University College London
London, United Kingdom
Amsterdam City Archives
Amsterdam, The Netherlands
cristina.casablancas.14@ucl.ac.uk

JOSEP GRAU-BOVÉ
University College London
London, United Kingdom
josep.grau.bove@ucl.ac.uk

TOM FEARN
University College London
London, United Kingdom
t.fearn@ucl.ac.uk

MATIJA STRLIČ*
University College London
London, United Kingdom
m.strlic@ucl.ac.uk

*Author for correspondence

KEYWORDS: collection surveys, damage function, libraries and archives, preventive conservation

ABSTRACT

Random collection surveys can be a rich source of data on the material state of a collection. However, random surveys do not necessarily provide data on the causes of degradation of collection items, which is useful in terms of resource allocation. For this, the reliability theory provides the required concepts. Using appropriate survey methods and statistical methods of data analysis, the so-obtained observational 'epidemiology' data has revealed risk factors that can lead to such degradation. Patterns in the observed data were identified that corroborated experimental research findings and enabled 'demographic' modelling of the dynamics of future change to be carried out in the surveyed collection for the case study of the Amsterdam City Archives. This study shows how, using appropriate methods of collection surveying, empirical and modelling studies of real collections can be successfully integrated, leading to useful evidence supporting collection care decision making.

INTRODUCTION

In preventive conservation, observational evidence plays an important role in decision making. The development of accessible monitoring tools and data mining techniques has enabled bigger and more detailed data to be collected than ever before; however, approaches are also being developed to maximise the utility of such data. In their previous research, the authors explored how archival and library users could inform collection care, specifically long-term planning horizons (Dillon et al. 2013) and the concept of 'fitness for use' (Strlič et al. 2015a). Using experimental approaches, models of fitness for use, i.e. damage functions, were developed in relation to wear and tear (Strlič et al. 2015b) and to chemical degradation (Strlič et al. 2015c). This work enables us to model change in populations of objects, and develop demographic curves for collections describing the 'survival' of objects in relation to the frequency of use, environmental parameters and material properties.

In addition to the development of models of fitness for use, a new approach to collection surveying was proposed to gather evidence on how mechanical degradation occurs and accumulates in the actual setting of an archival collection (Duran et al. 2017a). On the one hand, this approach is based on the principles of reliability, such as the failure rate and classification of failure causes, to decide which factors should be included due to their contribution to mechanical degradation (Duran et al. 2017b). On the other hand, using an epidemiological approach, groups of objects were analysed differing in their level of exposure, understood as the factors that might affect the degree of mechanical failure. In this work, the factors that have a greater effect on the event of mechanical failure and the patterns of decay in specific groups of objects were identified.

In this paper, the previous work conducted experimentally is combined with the results of an observational study conducted at the Amsterdam City Archives. Using the existing damage function for historic paper (Strlič et al. 2015c), the survey data was used to model the demographic curve describing the survival rate of the surveyed case study population. At the same time, responding to the growing interest in the collection of data to validate experimental results (van Duin 2014), the authors explored how epidemiological data can be used to validate the experimental model and therefore provide evidence to support decision making.

METHODOLOGY

The design and the analysis of the survey conducted in the Amsterdam City Archives have been described in detail elsewhere (Duran et al. 2017a). However, a brief description of the data used in this paper follows.

The survey sample consisted of 186 inventory numbers (stacks with loose sheets), from the 17th to the 21st century, selected on the basis of a good distribution of the outcome variable understood as mechanical failure, and factors that might affect the degree of failure, defined as exposure variables. The exposure variables were selected in order to obtain a good distribution during sampling: the number of requisitions in the reading room since 1998 (the year when reliable data became available as the result of the introduction of computers), the age of the document, protection, and the thickness of a stack. This is important as it enables the use of exploratory data analysis methods. However, this type of sampling is not calculated on the basis of a random survey and is thus not representative of the collection as a whole.

Throughout the study, data were collected and analysed on two separate levels: the inventory number as a whole (stack of sheets) and single sheets. Table 1 reports on how the outcome variables were defined related to the assessment of mechanical failure, and the factors that might affect the degree of failure as the exposure variables.

Each sheet (page) of an inventory number was categorised according to the failure category. Then, one single sheet for each failure category and one sheet with no signs of failure were randomly selected. This means that for each inventory number, 1 to 4 sheets were selected depending on whether or not a certain failure category was present within the inventory number. The total selection consisted of 431 paper sheets. 'Design' can be understood as the physical description of the object: size, weight, thickness of a volume or paper, i.e. any physical characteristic of an object that could affect its use. Only thickness (mm) was considered in this study – as shall be seen later, this affects the outcomes most strongly. Portfolios and boxes (i.e. protection) correspond to maintenance applied to objects.

Chemical and mechanical properties of paper were determined non-destructively using the SurveNIR instrument (Lichtblau e.K., Dresden, Germany), while instances of mechanical degradation (failure) were counted.

The data was analysed using MiniTab 17 statistical software. Exploratory data analysis methods were used, e.g. principal component analysis (PCA), as well as regression methods.

The damage function for historic paper (Strlič et al. 2015c) was used to interpret the data collected in the survey. The function expresses the rate of paper degradation dependent on the temperature and relative humidity during storage, as well as on the current pH and degree of polymerisation (DP) of paper. In addition, it calculates the number of instances of damage (represented by pieces of pages with text missing as a consequence of use), as a function of the instances of use by a reader under the conditions of general access in a reading room. A demography plot showing the proportion of the collection fit to be used was developed specifically

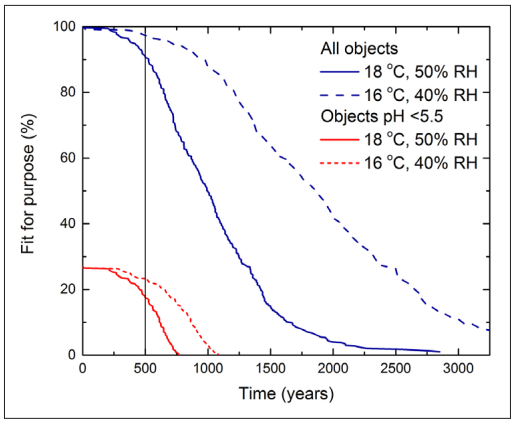

Figure 1. Demographic plots for all the surveyed objects, based on current pH and degree of polymerisation of the objects, as well as storage as indicated, with assumed average use of 0.5/year. Separate demographic curves for objects with pH < 5.5 are represented by (C) at 18°C and 50% RH, and (D) at 16°C and 40% RH. The vertical line represents the long-term planning horizon of 500 years

for the surveyed dataset, using four parameters: the current DP and pH values of the paper sample, the frequency of use and threshold of fitness for use. In this study, the strictest threshold of fitness for use was chosen (1 failed page/100 pages). The selection of frequency of use (0.5 instances of use/year) is discussed in the next section.

It needs to be stressed that one of the limitations of SurveNIR is that it cannot provide DP data for groundwood paper. This means that, in this case study, the demographic curves are representative of collections containing rag and bleached pulp paper only.

It is also important to note that in the applied damage function an object was considered failed only when a piece of text was missing (Strlič et al. 2015a). In contrast, the aim of the survey was to identify at what point mechanical degradation occurs, and therefore tears also counted as failure. In this paper, the plots presenting the survey data thus include all categories of failure. Although the same trends are obtained when only failure 3 (missing pieces, Table 1) is included, the trends are clearer when tears are added.

Table 1. Summary of the observational study design as used in the survey, including the outcome and exposure variables assessed according to two levels: inventory numbers (stack of sheets) and single sheets, and the data source

	VARIABLE		LEVEL	
			1. Inventory number (stack of sheets)	**2. Single sheet**
Outcome	Number failed pages (total)		Visual observation	---
	Number failed pages (failure 1): up to 5 tears, smaller than 2 cm		Visual observation	---
	Number failed pages (failure 2): more than 5 tears smaller than 2 cm and/or tear(s) larger than 2 cm		Visual observation	---
	Number failed pages (failure 3): when a piece was missing, regardless of the number of tears		Visual observation	---
	Failure type (1-4)		---	Visual observation
Exposure	Design	Thickness (mm)	Physical measurement	Physical measurement
	Maintenance	Protection (portfolio or box)	Visual observation	---
	Usage	Number requisitions in reading room since 1998	Collection management system	Collection management system
		Date (age)	Collection management system	Collection management system
	Manufacture	pH	---	Near infrared spectrometry
		Degree of polymerisation	---	

RESULTS AND DISCUSSION

In Figure 1, the demographic plot showing the percentage of objects fit for purpose (here defined as containing 1 failed page per 100 pages), is shown. This assumes storage at 18°C and 50% relative humidity (RH) and 16°C and 40% RH in average, which enables us to explore how a small change in the environmental conditions might affect the future fitness

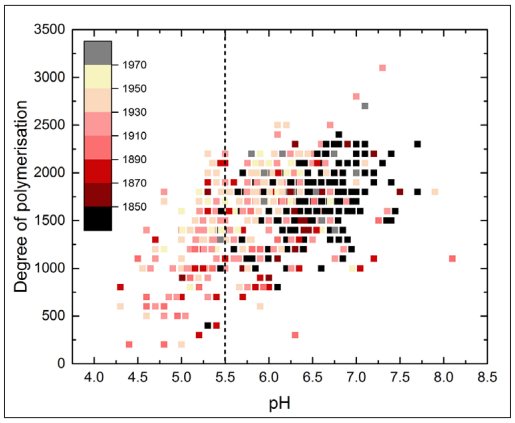

Figure 2. Scatterplot showing the degree of polymerisation and pH values for the surveyed paper. The dashed vertical line separates objects with pH < 5.5 (cf. Figure 1)

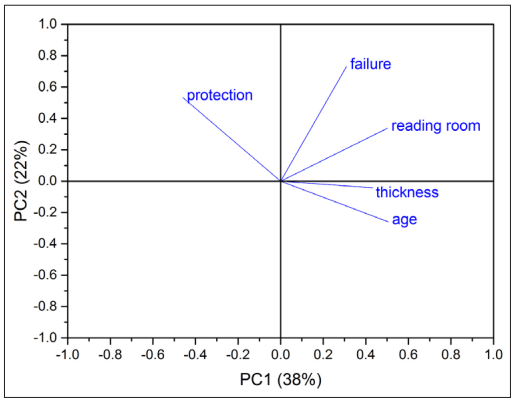

Figure 3. PCA loadings, showing correlations between failure (percentage of failed pages within an inventory number), number of requests in the reading room, type of protective enclosure, age and thickness of inventory numbers with loose sheets. A factor that strongly correlates with accumulation of failure needs to be parallel to the loading for 'failure' – in the above plot this was the number of instances in which the object was delivered to the reading room. The % variance explained by the principal components is in parentheses

for use of the surveyed collections. In addition, the plots are based on one instance of reading every two years, which represents a rather high average frequency of use as will be shown later. In the calculation of the demographic curves, pH and DP data were used to calculate the expected time until failure for each single object used in the survey. This is defined by both gradual chemical degradation due to the degradation of cellulose as the most important structural macromolecule in paper, and by mechanical stress as induced during instances of physical use.

In a previous study (Strlič et al. 2015c), it was generally observed that two waves of degradation could be seen when the demographic curve is plotted for the whole of the population: the first wave representing the acidic paper and the second one representing the rest of the collection. In contrast, the curves reflecting the entire surveyed collection (Figure 1) show only one wave of degradation.

This difference might be explained by analysing the acidic paper group only. Of the assessed sheets of paper, 27% showed pH < 5.5 (Figure 1), and only a very small proportion of 7% had pH < 5. Although loss of fitness of this group begins at a slow rate, it increases fast as the long-term planning horizon approaches (Figure 1, vertical line), as defined in a recent study of attitudes towards damage in collections, in which it was found that ~90% of library and archival users would find a 500-years long-term planning horizon acceptable (Dillon et al. 2013).

According to the demographic plot, loss of fitness could be more than halved (from 36% to 14% for acidic paper, and from 10% to 2% for the surveyed collection as a whole) in the next 500 years by slightly lowering the temperature and relative humidity. These results validate the approach to management of collections as advocated by the recent performance-based collection care standards (PAS 198:2012).

The current chemical state of the group of acidic papers is shown in Figure 2; the group of objects more at risk are mostly papers dating between 1870 and 1950. These results are corroborated by the survey data that showed a clear difference in the proportion of mechanical failure between less and more accessed documents dating after 1850 (Duran et al. 2017a). In addition, the group of inventory numbers with a low frequency of use showed a consistently lower rate of deterioration, i.e. accumulation of mechanical degradation, than those that are accessed more frequently.

The importance of instances of physical use in the damage function used in the demographic plot becomes clearer when the survey data is further analysed. Not entirely surprisingly, principal component analysis (PCA) of the variables shows that the percentage of failed pages (including tears and missing pieces) is strongly correlated with the number of requests (Figure 3). With 8 requests since 1998 as the threshold, the mean percentage of failed pages by inventory numbers requested less than 8 times is 18% (*SD* 22.2, *SE* 3.1), while the mean for highly accessed records is 36% (*SD* 25.6, *SE* 2.2). According to the two-sample *t*-test, this difference is statistically significant (95% *CI* for *M* difference -25.1, -9.9, t=-4.58, df=100, p <0.001) and indicates that use in the reading room has a significant effect on failure.

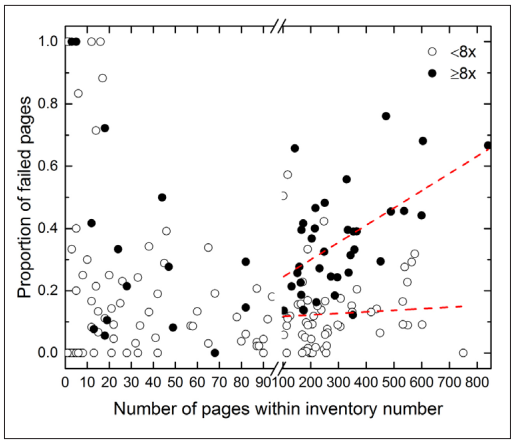

Figure 4. Scatterplot showing the percentage of failed pages of inventory numbers depending on the number of times that the number has been accessed in the reading room since 1998. Note that the data is presented at different scales for less than and more than 100 loose pages to accentuate the observed differences. The dashed linear regression lines are for observations corresponding to >100 pages only

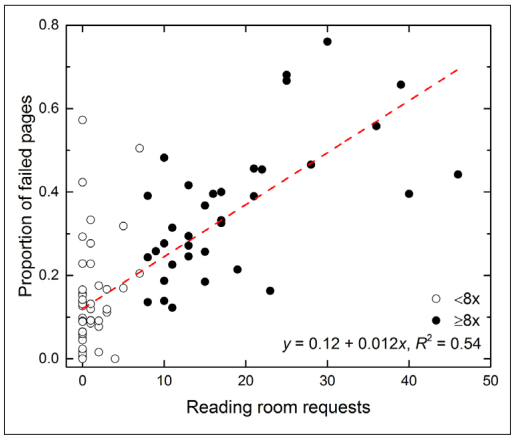

Figure 5. Proportion of failed pages within an inventory number depending on the number of times that the number has been accessed in the reading room since 1998. The plot shows inventory numbers containing more than 100 loose sheets (pages) only. The regression line is drawn using all data points regardless of access

In the PCA loading plot, loadings for variables that are closely related are parallel, while loadings for variables that are less related are perpendicular to each other. Thus, Figure 3 also indicates that thicker and older inventory numbers are less frequently protected, and that there is no or very weak correlation between failure and protection, and age and thickness of an inventory number. However, it shall be seen that these correlations can be further explored using linear regression.

An important aspect for the accumulation of wear and tear, as well as for the validation of the demographic plots, is the stack thickness. The damage function used for the demographic plots has been developed in experimental studies modelling stacks of 50 sheets (Strlič et al. 2015b). Corroborating the results of these experimental studies, data collected during the survey show that for inventory numbers containing less than 100 sheets with an average DP higher than 800, no clear pattern of failure can be seen (Figure 4). But the same data also indicate that the thickness of an inventory number affects the accumulation of mechanical damage when inventory numbers are thicker than 2.5 cm (equivalent to inventory numbers containing more than 100 sheets). The relevance of stack thickness is further stressed by the good association between the percentage of failed pages in an inventory number containing more than 100 sheets and the number of requests (Figure 5).

Figure 5 shows at what point wear and tear accumulates as a function of the number of accesses in the reading room, and institutions could use this plot to support their preservation policies and to determine at what point handling in the reading room would result in an unacceptable level of damage. In the case of the Amsterdam City Archives, when records have been requested more than 8 times, no data points can be found with less than 10% of failed pages (which includes tears and missing pieces).

The Amsterdam City Archives has been recording the number of requests since 1998 and the survey took place in 2016. It is thus possible to calculate how many times in average an inventory number has been requested per year during this period. For example, the inventory numbers that have been accessed more than 8 times since 1998 have been requested ~0.5 times a year in average. This threshold has thus been used to determine the parameter of frequency of use to develop the demographic plot.

It has been shown that, according to the survey data, failure predominantly depends on use, which has been experimentally determined before; however, the effect of material type and thickness of an inventory number has now also been shown. These findings support the use of demographic plots to examine the effect of storage conditions on the loss of degree of polymerisation, which is related to loss of material strength and thus accumulation of mechanical failure as a consequence of use (Strlič et al. 2015b).

On this basis, in the new repository of the Amsterdam City Archives (mostly containing paper dating after 1850) the temperature will be lowered to 16°C. At the same time, and as a result of the survey, a project will be started to digitise the most frequently requested items in the reading room,

the current suggestion being to start with those that have been requested more than 20 times since 1998.

CONCLUSION

This paper shows, using the example of a paper-based archival collection, how observational epidemiology data can be successfully integrated with demographic modelling of the dynamics of change in collections. When surveys are performed in agreement with the requirements of reliability engineering, in which case causal relationships between failure and exposure variables can be uncovered using multivariate and univariate data analysis techniques, then such approach provides significantly more detail and insight into the mechanisms of failure in archival objects, as well as validates the experimental model of accumulation of mechanical deterioration during instances of access.

Using the collected material properties (pH and degree of polymerisation) in conjunction with the observed frequency of use, a demographic plot for the case study collection has been developed. Since the rate of mechanical degradation increases significantly once the degree of polymerisation is less than 800, the most significant prevention measure is still environmental control; in the long-term planning horizon, loss of fitness could be halved by decreasing storage temperature by 2°C and relative humidity by 10%.

As the proportion of failed pages increases linearly with the number of requests, the failure rate for inventory numbers with loose sheets can be determined as a function of the number of times accessed. While this is perhaps an expected outcome, important observations emerged about the material type, inventory number thickness, protection and age that influence the outcome. The study has important practical implications for preventive conservation of archival documents, specifically in relation to the thickness of archival folders, which have shown to have a significant effect on accumulation of random failure. Furthermore, it shows how the effects of frequent physical use could be counterbalanced by access to digitised copies, as well as how much failure could be ascribed to the process of digitisation, were this to take place at regular intervals in the future.

The developed demography plots show that in relation to how archival and library collections are currently valued by users, the Amsterdam City Archives could be preserved in a manner that meets their expectations.

ACKNOWLEDGEMENTS

The support of the Amsterdam City Archives throughout the project is gratefully acknowledged. The authors are also grateful to the following colleagues from the Archives: Erik Schmitz, Jochem Kamps, Janien Kemp and Emmy Ferbeek for comments, to OmniAccess for conducting the measurements with SurveNIR, and Prins Bernhard Cultuurfonds for the financial support. This work was carried out in the frame of the MRes Science and Engineering in Arts Heritage and Archaeology (SEAHA), Institute for Sustainable Heritage, University College London.

REFERENCES

DILLON, C., W. LINDSAY, J. TAYLOR, K. FOUSEKI, N. BELL, and M. STRLIČ. 2013. Collections demography: Stakeholders' views on the lifetime of collections. In *Climate for Collections Conference, Munich, 7–9 November 2012*, eds. J. Ashley-Smith, A. Burmester, and M. Eibl, 45–58. London: Archetype.

DURAN, C., J. GRAU-BOVÉ, T. FEARN, and M. STRLIČ. 2017a. Epidemiology for the study of mechanical degradation in archival collections: A methodology study. Submitted for publication in *Heritage Science*.

DURAN, C. and M. STRLIČ. 2017b. Wear and tear in archive and library collections: How does it happen? Submitted for publication in *Studies in Conservation*.

PAS 198:2012. 2012. *Specification for managing environmental conditions for cultural collections*. BSI: London.

STRLIČ, M., C. GROSSI-SAMPEDRO, C. DILLON, N. BELL, K. FOUSEKI, P. BRIMBLECOMBE, E. MENART, K. NTANOS, W. LINDSAY, D. THICKETT, F. FRANCE, and G. DE BRUIN. 2015(a). Damage function for historic paper. Part I: Fitness for use. *Heritage Science* 3: 33.

STRLIČ, M., C. GROSSI-SAMPEDRO, C. DILLON, N. BELL, K. FOUSEKI, P. BRIMBLECOMBE, E. MENART, K. NTANOS, W. LINDSAY, D. THICKETT, F. FRANCE, and G. DE BRUIN. 2015(b). Damage function for historic paper. Part II: Wear and tear. *Heritage Science* 3: 36.

STRLIČ, M., C. GROSSI-SAMPEDRO, C. DILLON, N. BELL, K. FOUSEKI, P. BRIMBLECOMBE, E. MENART, K. NTANOS, W. LINDSAY, D. THICKETT, F. FRANCE, and G. DE BRUIN. 2015(c). Damage function for historic paper. Part III: Isochrones and demography of collections. *Heritage Science* 3: 40.

VAN DUIN, P. 2014. Climate effects on museum objects: The need for monitoring and analysis. *GCI Newsletter* 29(2): 13–5.

How to cite this article:
Duran, C., J. Grau-Bové, T. Fearn, and M. Strlič. 2017. Data mining in collections: From epidemiology to demography. *In ICOM-CC 18th Triennial Conference Preprints, Copenhagen, 4–8 September 2017,* ed. J. Bridgland, art. 1516. Paris: International Council of Museums.