

Manuscript Boxing

A Technique for Objective Spatial Arrangement

Books from historic library collections often require protective boxes which typically increase the required storage space. In some cases, extra storage space is unavailable and therefore good use of existing space is necessary. Deciding on the location of every boxed book on the available shelving is not an easy problem. Calculating the dimensions of the boxes when several hundreds or even thousands of books are considered is again time-consuming. Both of these problems are extremely difficult to solve manually. There are several techniques from the fields of statistics and computer science which can be used to assist with finding a solution. This paper investigates k-means clustering as a way to group books which can fit in a box of a specific size. It also investigates the First Fit Decreasing technique for arranging boxes optimally on shelves. The paper also includes a brief description of the boxes used at the library of the St. Catherine Monastery in Sinai, Egypt, alongside the racking storage system. It describes an application implementing the above techniques for the library's boxed manuscripts and presents some results confirming the usefulness of the application.

Aufbewahrung von Manuskripten in Schachteln: Eine Methode zur zielgerichteten räumlichen Anordnung

Bücher historischer Bibliotheksbestände benötigen oft schützende Schachteln, die den zu ihrer Lagerung benötigten Platz vergrößern. Mitunter ist zusätzlicher Lagerplatz jedoch nicht verfügbar, was eine effektive Nutzung vorhandener Flächen erfordert. Die optimale Positionswahl für jede Buchschachtel auf der verfügbaren Regalfläche ist kein einfaches Problem. Eine Berechnung der Schachtelgröße ist zeitraubend, wenn es mehrere hundert oder tausende Bücher betrifft. Manuell sind beide Probleme nur schwer lösbar. Statistik und Computerwissenschaften bieten Techniken, die bei der Lösung des Problems helfen können. Die vorliegende Arbeit untersucht das k-Mittelwert-Clustering, um Bücher für spezifische Schachtelgrößen zu gruppieren, sowie die 'first-fit decreasing'-Methode zur optimierten Anordnung von Schachteln auf Regalflächen. Kurz vorgestellt werden außerdem die in der Bibliothek des Klosters St. Katharina (Sinai, Ägypten) verwendeten Buchschachteln, die Regale für deren Lagerung, die Umsetzung der genannten Methoden zur Aufbewahrung der Manuskripte dieser Bibliothek sowie einige Ergebnisse, die die Nützlichkeit dieser Anwendung aufzeigen.

The storage of library collections is an important logistical problem for librarians and conservators. When books require protection, the addition of protective boxes increases the amount of shelving space occupied and therefore careful space management is necessary. Calculating a good spatial arrangement of book boxes, to ensure that space is being used fully and all the boxed books fit in the available space, is not a trivial problem. Techniques borrowed from computer science and statistics can be used to help with the solution. In computer science, the bin packing problem poses similar questions and a range of solutions are available. Another interesting problem is grouping the manuscripts according to size. This is useful when there is a limit to the number of different box sizes available and when a choice of the best sizes is required. In statistics, cluster analysis addresses similar questions. Although these problems are both presented in this paper, it is important to emphasise that the techniques examined for bin packing and cluster analysis are independent. If, therefore, there is no limitation in the number of box sizes, box clustering is not required. Similarly, if there is no limitation in space, storage through bin packing algorithms is not necessary. Our approach allows any combination of techniques used according to parameters which can be set on demand.

This work was initiated during our study of the manuscripts at the St. Catherine's Monastery in Sinai, Egypt.

St. Catherine's Library and Condition Survey

The St. Catherine's Monastery in the Sinai peninsula is one of the earliest Christian monasteries in the world and can claim continuous monastic life from the 4th century. Previously

published work about the Monastery (Brock 1996; Manaphes 1990) has showcased one of the most significant collections of manuscripts in the world. Most of these date from before the 16th century and maintain their original or early binding structures, making the collection a valuable resource for the study of the history of bookbinding. The importance of the material led to an extensive condition survey of the collection which started in 2001 (Pickwoad 2004). The survey was undertaken by Camberwell College of Arts (University of the Arts London), during which the condition of the bindings was assessed and recorded in a database (Velios and Pickwoad 2005). To ensure the protection of some manuscripts it was considered necessary to box them and store them horizontally. It was agreed that the books do not need to be stored in shelfmark sequence and that a finding aid is used to locate them on the shelves. A combination of visual observations and a set of database queries were used to decide which books should be boxed. The current number of manuscripts in need of a protective box is about 2,000, the rest will remain vertically shelved. A separate project was initiated to design and specify a prototype box to be used at the library as described in the following section.

At the same time, a large development project has been proposed for the library at St. Catherine's. The weak structure of the library building poses risks for the collection in the event of an earthquake and therefore plans for rebuilding the library rooms are being discussed. In the new library building, the boxed manuscripts will be housed in alcoves set in the library walls. An example of an alcove is shown in Fig 1. There are 31 such alcoves in the new library design, each with known dimensions. The detailed condition survey also means that information

about the size of each book is available and the objective of the work presented here is to establish whether boxed books can fit in the available alcove spaces.

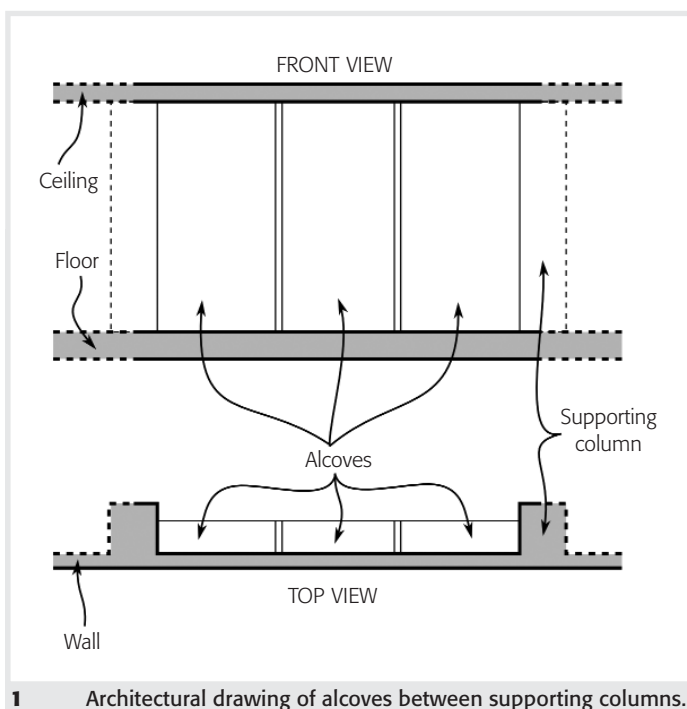
The Problem of Objective Spatial Arrangement

Traditionally, similar problems in conservation have been dealt with by experienced conservators or technicians. Experience would allow a conservator to assess the extra size of the boxing material around the books for the whole collection and give a reasonable estimate on whether extra shelving is needed or whether the current shelving is adequate. Hadgraft (Hadgraft 1994) described a project with similar objectives where a process using complex calculations was employed. This relied on a manual trial and error approach to identify optimum parameters but the results demonstrated the benefits of using such calculations. In our case it was difficult to estimate the shelving size based on experience because it was the first time that this design of box and stack was being used. A brief description of this design is included in the next paragraph. Also our intention was to identify objective criteria for the arrangement of the boxed manuscripts rather than a subjective trial and error approach.

Box and Stack Design

Previous work undertaken by the Ligatus Research Centre [1] in collaboration with Conservation by Design, has resulted in a prototype for a steel book-box (Fig 2). The unusual choice of material will be explained extensively in Pickwood and Honey (Pickwood and Honey in preparation) and can be summarised here:

- > In extremely dry conditions, such as those of the Sinai desert, traditional book box materials tend to deform, whereas steel is not affected by low humidity and therefore retains its original shape.
- > Precision manufacturing techniques borrowed from the steel industry allow the semi-automatic manufacturing of boxes



more quickly and precisely than manufacturing techniques for traditional materials.

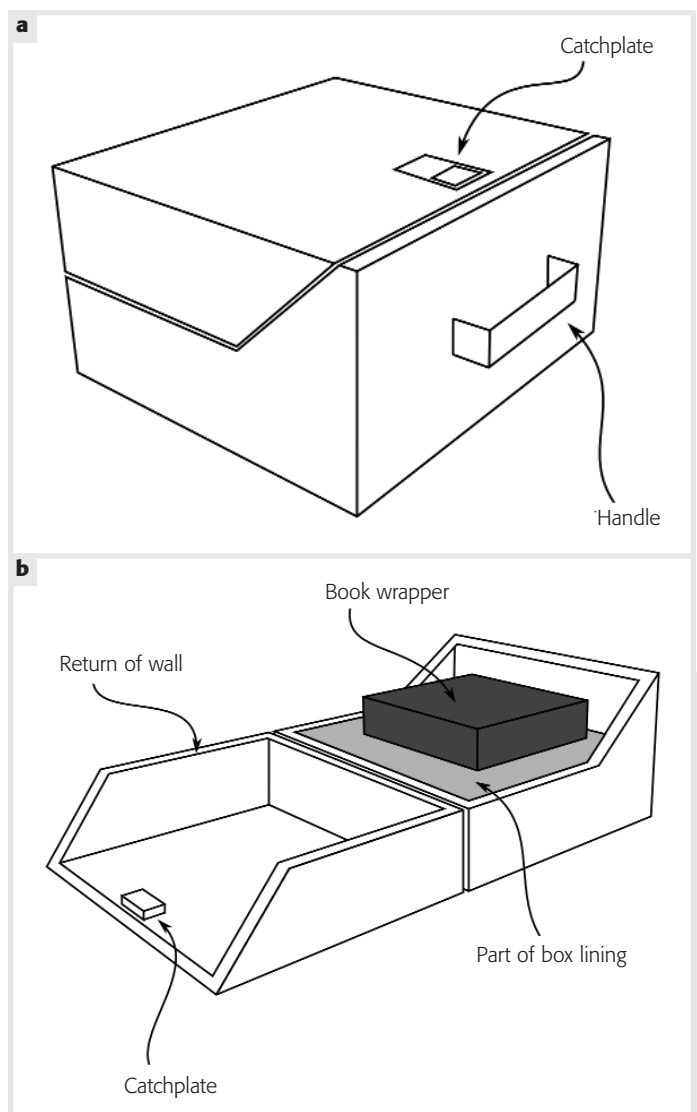
- > Although the cost of manufacturing a steel box is somewhat higher than the cost of manufacturing a tailored buckram-covered drop spine box, its superior strength and durability make it more than competitive in price.

Overall Box Description

The box is of parallelepiped shape with a protruding handle (Fig 2a). The book is oriented in the box so that the book and box dimensions align as listed in Tab 1 and shown in Fig 2b [2].

The manufacturing process dictates a minimum size limit to the box in each dimension (Tab 2). None of the books in the collection is large enough to give concern about maximum dimension limitations.

The box has been designed in such a way that internal linings of variable thickness can be slotted into place to support the wrapper containing the book. The return of the side walls of the box under which the lining is slotted is about 20 mm wide (Fig 2b). A maximum limit of 200 mm of lining material in any



Tab 1 Orientation of a book inside a box.

Book dimensions	Box dimensions
Thickness	Height
Height	Width
Width	Depth

one dimension (100 mm on either side of the book) was set to avoid possible movement of the lining material from its original position. Tab 2 shows the minimum padding sizes for each box dimension. These figures include the thickness of the steel sheet material and the thickness of the wrapper material.

Although in theory each box can be manufactured using a different set of dimensions, in practice a fixed number of different sizes (e.g. 20 different sizes of boxes) could prove more economical. The largely automated manufacturing process still involves some manual work. The requirement that the lid and base are of equal height and the consequent variations in the depth of the sloping leading edge of the two trays to the overall box height mean that the manufacturer would prefer to limit the number of different box sizes so that the manual stages of the manufacturing become more efficient. Repeating the same process for multiple boxes is faster than revising it for each box.

Stack Description

The boxes are stored flat on shelves suspended on purpose-built racks (Fig 3). Each rack is mounted on the wall and it includes slots drilled at regular intervals of 20 mm, from which the shelves are suspended. Drilling holes at fixed intervals means that the racks are standardised and can be produced automatically by a machine. Because the racks are parallel, all of the mount-

Tab 2 Box dimensions.

Dimensions [mm]	Min box size	Min padding size	Min margin
Height	100	40	10
Width	200	38	10
Depth	150	150	33

ed shelves on two racks must have the same width. Therefore, and in order to avoid movement of the boxes while stored on the shelves, all boxes held on a pair of racks must have the same width.

Boxes are positioned on separate shelves so that they can be removed without other boxes being disturbed. A margin between the different shelves is set (height) to allow the removal of a box. The margin for the width is minimal whereas the margin is more generous for the depth in order to accommodate the handles. The margin figures are shown in Tab 2.

Large boxes containing heavy books can be difficult to handle and placing the largest boxes in the middle of each vertical stack makes sense from an ergonomics point of view, as lifting heavy weights at waist height is safer than from low or high stack positions.

The above figures and requirements have been set either by design or by the manufacturer and were important in the consideration of the automated process for calculating the box dimensions and their arrangement on the available space.

Objective of the Project

Having received the specifications for the box and the stacks, our objective was to develop a process which would begin with inputting the dimensions of the books and alcoves and would conclude with a list of box sizes with their position on the stacks.

Clustering and Bin Packing Methodologies

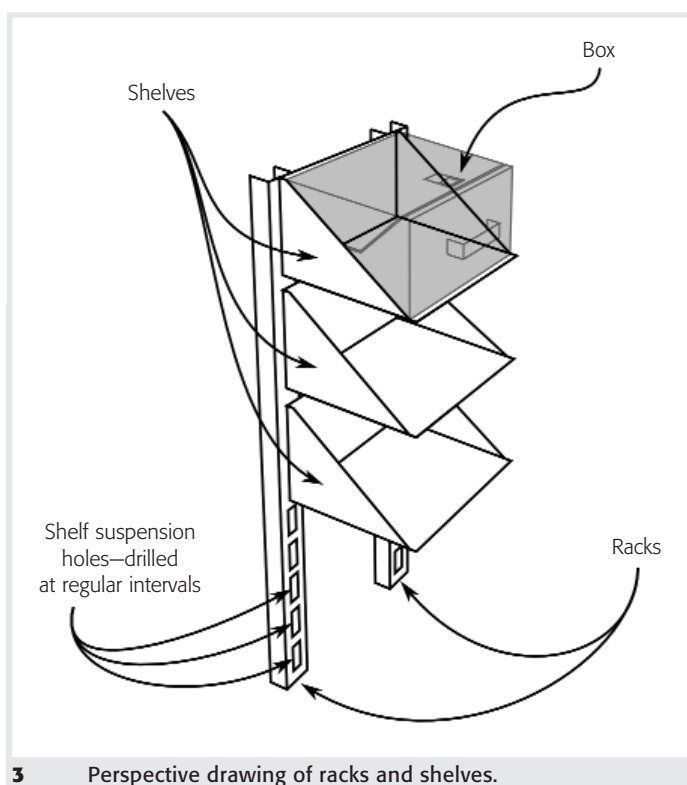
Box Size Calculation

Calculating the minimum dimensions for each box is simple. Adding the minimum lining to the overall dimensions of a book gives the resulting dimensions of the corresponding box (e.g. book height + width lining = box width). In order to ensure that no box is smaller than the minimum box dimensions, a comparison of each calculated box dimension to the figures of minimum box size (Tab 2) shows whether the calculated box dimension is too small. In which case the figure increases to the minimum allowed. This is done for each book and gives an initial list of box dimensions according to the specification.

Box Arrangement

The process described in the previous section makes no provision for the possibility of a fixed number of box sizes (to allow for more economical manufacturing) and assumes that every box is unique. Grouping the boxes according to size but still ensuring that all books fit in a box is a complex problem, also known as clustering.

The box dimensions resulting from the process described in the previous section are ordered according to the shelfmark of

**3** Perspective drawing of racks and shelves.

the book (i.e. the original order of the books in the library). That order does not take into account optimisations for saving space or placing the books on shelves of the same width as the stack design requires. Finding a good (if not the best) order in which these boxes can be arranged to save most space is not trivial and is also known as the bin packing problem.

In the following section we investigate these two problems and suggest possible solutions and their application to library collections.

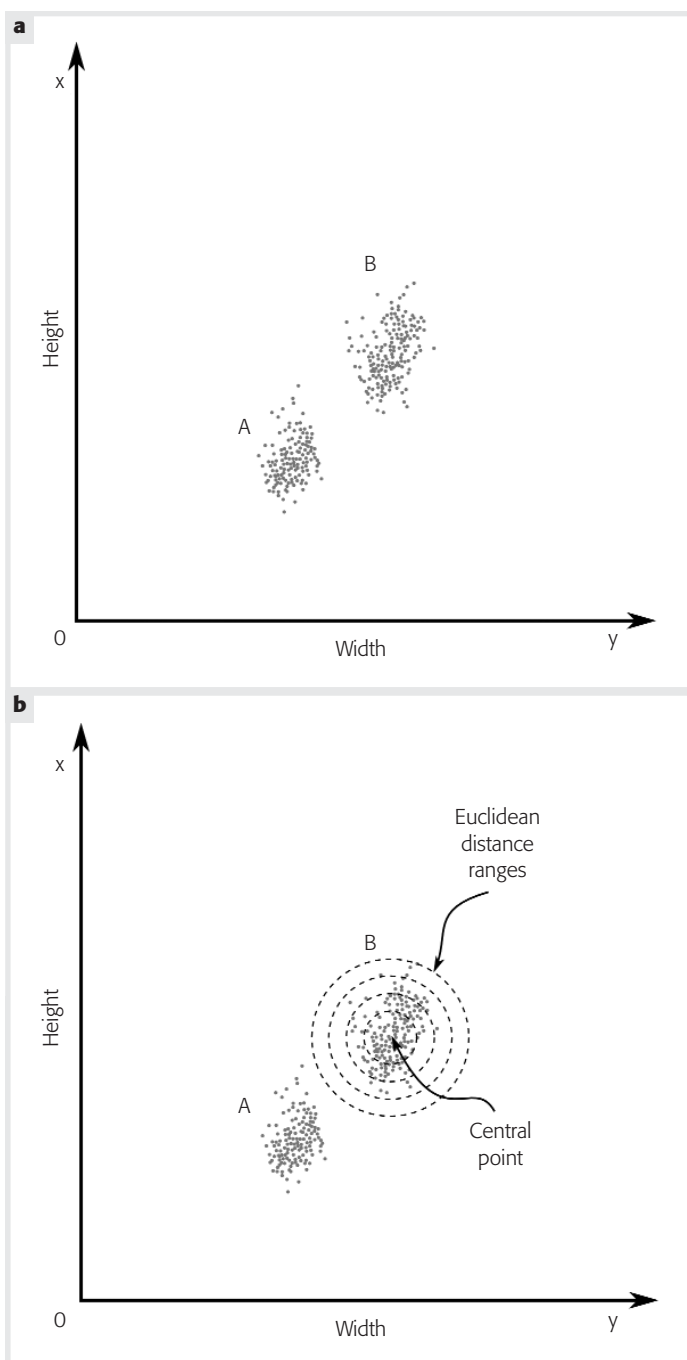
Clustering

Clustering could be explained as the division of a set of measurements into groups based on similarity. For the problem of pro-

ducing a fixed number of box sizes (e.g. 20) for about 2,000 books with different dimensions, clustering describes the process of grouping these books according to their dimensions into 20 groups. For example, any book in group 1 should fit inside any box of size 1, any book in group 2 should fit inside any box of size 2, etc. Similarity is assessed using a so-called metric criterion which is better understood if measurements are plotted on a graph. For example in Fig 4a, we have plotted the sizes of some books where the X axis corresponds to the height of the book and the Y axis corresponds to the width. There are two groups (clusters) visible in that chart. A popular way to indicate this mathematically is by calculating the Euclidean distance of every point from a central point (Fig 4b). A short distance indicates that a point is close to the central point of a cluster and therefore is part of it, whereas a long distance indicates that a point is far away from the cluster and therefore it probably belongs to a different one. In some cases other types of distance can be used to measure similarity of data. For example, the Mahalanobis distance is another metric which takes into account the variance of the data and therefore it is suitable for measuring similarity in non-uniform data (De Maesschalck et al 2000). However, the Euclidean distance remains a popular (if not the most popular) metric when measuring spatial data and therefore we decided to adopt it here.

A number of different techniques have been identified to solve clustering problems. Jain et al (Jain et al 1999) describe these techniques and the advantages for each. Their review is useful for choosing a technique to apply to our box size problem. They divide clustering techniques into hierarchical and partitional. Visualising hierarchical clusters typically results in a figure similar to what is shown in Fig 5. The principle is that each point in the whole dataset can be considered as one cluster which can be merged according to a criterion (and with the Euclidean distance as a metric) with a neighbouring cluster. Each one of these new clusters can be further merged with neighbouring ones and so on, until there is only one cluster left which corresponds to the whole dataset. Clustering based on hierarchies is suitable for applications of automatic classification, where a point can be traced back to its ancestor and therefore it can be said that it belongs to this ancestor's class. Because of this overall indication of the classification of data, hierarchical clustering is suitable for inter-relating data (for example Gauch and Whittaker 1981) but it does not produce a fixed number of clusters directly, as discussed, among others, in Sander et al (Sander et al 2003). A line could be drawn, as shown in Fig 5, which could limit the hierarchy to a desirable number of clusters but it is difficult to justify where this line should be. In practice our book sizes dataset was examined with the popular tool for cluster analysis R (Gentleman and Ihaka 2010) which utilises hierarchical clustering techniques. This confirmed that the resulting solutions were not practical in grouping the books. Because hierarchical clustering techniques cannot directly indicate a definitive grouping to our box size problem, partitional clustering was considered.

Partitional clustering does not rely on relations among data. It looks at data and separates it. In their review Jain et al (Jain



4 Euclidean distance as a metric for clustering: sizes of books indicate two clusters (a); Euclidean distance from a central point separates one cluster from another (b).

et al 1999) emphasise a critical problem in partitional clustering: deciding the number of clusters before the dataset is divided. This is a difficult issue in many applications of clustering, but in our case the box manufacturer decides on the number of box sizes and therefore the number of our clusters is known in advance. A typical partitional clustering solution can be described as follows:

- > The division of the dataset starts from central points (as many as the desirable number of clusters). There are many ways to decide the initial central points and one of them is discussed later in this section.
- > All points are assigned to clusters according to a calculated metric (e.g. Euclidean distance) and a criterion for including or not including a point to a cluster.
- > Following that, the central points are reassigned to better reflect the cluster's position and the metrics for each point recalculated.
- > This process is repeated until all points remain in the assigned cluster after a new cycle or the differences in the metrics calculated is insignificant.

Various partitional clustering algorithms are available and evaluating them and choosing one is a challenging research problem beyond the scope of this paper. Existing literature points to a popular partitional technique which is called the k-means solution and was originally proposed by MacQueen (MacQueen 1967), but its value has been confirmed more recently (in computer science Chen 2005; Jain 2010; and, for example, in medicine Ubeyli and Dogdu 2010). The k-means algorithm follows the steps described above with the initial central points chosen at random. It makes use of the so-called squared error criterion which means that a good set of clusters has been identified when the sum of the distances, from each point of every cluster to its central point is the minimum. In other words, the central point for every cluster is positioned in such a way that it results in the shortest distances from all points in its cluster [3]. Implementa-

tions of the k-means algorithm are relatively efficient and therefore k-means is also a good choice in terms of speed.

The k-means solution can work for data in multiple dimensions. Typically for spatial arrangement of boxes a three-dimensional solution is required. In the case of the St. Catherine's new library a two-dimensional solution has been used to cluster the boxes in width and depth as explained in the section about 'Implementation and Results'.

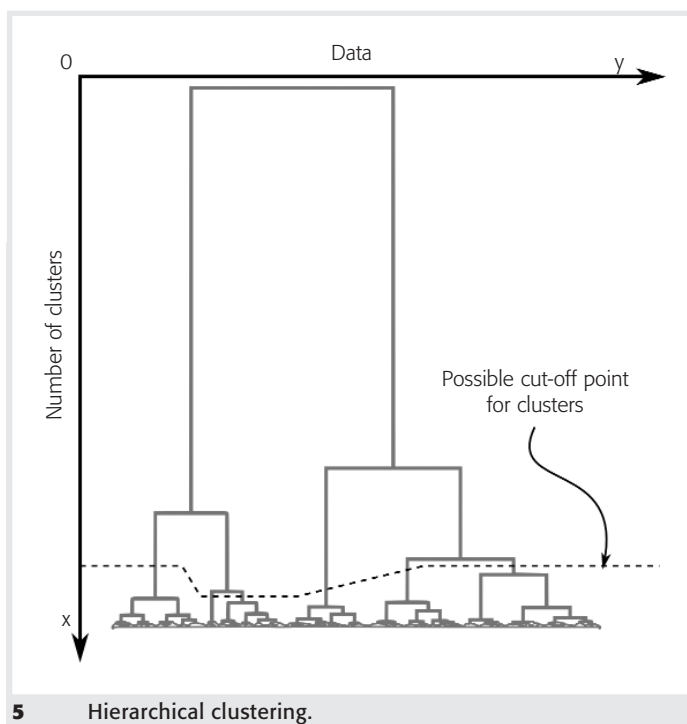
Bin Packing

The problem of bin packing is described, among others, by Korf (Korf 2003): Given a set of numbers (in our case box heights and widths), and a set of bins of fixed capacity (in our case the new library's book stacks and alcoves), the NP-complete problem of bin packing is to find the minimum number of bins needed to contain the numbers, such that the sum of the numbers assigned to each bin does not exceed the bin capacity. In other words, what is the minimum number of stacks needed to store all of the boxes? NP-complete is a term used in computational theory and in plain language it means that a problem is impossible (or perhaps extremely difficult) to solve in realistic time [4]. As such, this problem can be solved by following a range of different techniques but the solution recommended may not be the best. In practice these techniques offer good enough solutions for everyday problems, as will be explained in the following sections.

The bin packing problem is often explored in two or three dimensions in typical logistical applications, i.e. a shipping company will look into fitting a group of three-dimensional parcels into a three-dimensional container or a manufacturing company will look into cutting a template from sheet material by fitting two-dimensional shapes into a two-dimensional rectangular sheet. In the case of the stacked boxes the problem is reduced to one dimension since boxes on a stack all have the same width, therefore the stack is packed along one dimension only (height). Once all stacks have been filled with boxes, a second bin packing problem is explored: fitting the stacks into the available width of the alcoves. Again this is a one-dimension bin packing problem since all stacks have the same height and therefore the space is filled across one dimension (width). The depth of the stacks does not concern us in this case as there is only ever one book on a shelf.

Bin Packing Algorithms

Coffman et al (Coffman et al 1996) give a good introduction to the various techniques used for bin packing. A theoretical problem is shown in Fig 6a where the given items need to be rearranged to occupy less space or fewer containers (we assume that the width of both units and containers is the same). A common solution is given by the Next Fit algorithm (Fig 6b) where the next item in sequence is tested on whether it fits in the current container. If it does, then it is packed there. If it does not, then a new container is used and the process moves to the next item. This is an extremely fast way for packing and works well when the items can only be examined sequentially, i.e. we have no idea of the dimensions of the following items. A practical application of this technique is packing truck containers, where holding



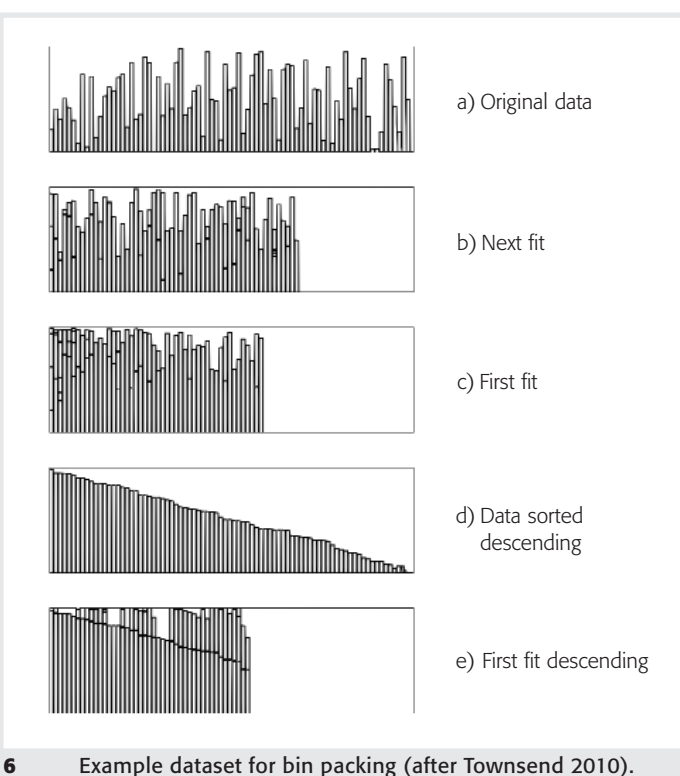
back a truck until its loading space is totally full is impossible because fast delivery is important. Book box packing, of course, has no such limitation.

Results are significantly better when using the First Fit algorithm (Fig 6c). In this case the next item in the sequence is tested on whether it fits in any of the existing containers which are not full. A new container is used only after an item does not fit anywhere else. Because the test of fit takes place for a larger number of possible containers, this algorithm packs items tighter than the Next Fit one. Results tend to improve when items are sorted by size (largest first) prior to the First Fit algorithm. The larger items occupy as much space as they have to and then the smaller ones follow to fill in gaps left in the containers. This is also known as the First Fit Decreasing algorithm (Fig 6e).

This simple example illustrates the advantage of the First Fit Decreasing algorithm compared to the other solutions. Many variations of the First Fit exist, but a thorough comparison of these is beyond the scope of this paper and extensive work on the subject is being done continuously elsewhere. The basic bibliography in the field (for example Skiena 1998: §8.6.9 or Coffman et al 1996: §2.2.7) indicates that the First Fit Decreasing solution is a popular and successful one which also offered satisfactory results in our case, as discussed later.

Implementation and Results

As mentioned previously, standard-size boxes reduce the cost of manufacturing but result in wasted space as the smaller books in a cluster still require a large box. Therefore a reasonable compromise needs to be made by increasing the cost to reduce wasted space. This discussion is beyond the scope of this paper, but it is, however, important to emphasise that clustering can be limited to only two dimensions (i.e. box width and box depth) thus



Tab 3 Cluster populations and max book sizes per cluster.

Cluster	Box count	Max book width [mm]	Max book height [mm]
0	178	246	322
1	116	164	219
2	205	167	244
3	56	349	423
4	17	198	192
5	58	148	159
6	105	170	203
7	74	138	187
8	98	150	225
9	99	281	370
10	50	157	172
11	18	106	133
12	161	235	298
13	100	129	177
14	115	204	225
15	159	184	257
16	148	201	286
17	73	214	253
18	67	118	156
19	105	155	189

allowing more granularity in the box sizes for height, which is a critical dimension since the boxes are arranged one above the other. In the case of the St. Catherine's new library, a quick look at the data will indicate that the width of the books (the depth of the boxes) is in general smaller than the available racking depth and a choice of fixed depths is reasonable for reducing manufacturing costs. The height of the books (width of boxes) is important because boxes are stored in groups on the racks according to width. Therefore applying the clustering technique to the height of the books to produce a fixed number of box widths is reasonable.

Finally, perhaps the most important dimension of the books for saving space is thickness (height of boxes), since the books are stacked along that dimension. A reasonable decision is to accept a higher cost and remove this dimension from the clustering process so that each box has an individual and minimum possible height on the rack. Our implementation accommodates such combinations on demand.

Box Size Calculation

In order to calculate the box dimensions for each book the k-means solution is applied to the book dimensions dataset. A popular data analysis tool, WEKA (Hall et al 2009), can be used for that purpose. The number of clusters for the data to be divided is specified by the manufacturer to 20. The clustering results, as shown in Fig 7, assign a cluster number to every book in the library. The results are summarised in Tab 3. The box width and depth for each cluster is calculated after adding the largest book height and width Tab 3 to the lining sizes shown in Tab 2. The height of the boxes is calculated for each box individually after clustering. Popular spreadsheet software can be used for that.

Having calculated the box dimensions the next step is their arrangement on the stacks.

Box Arrangement

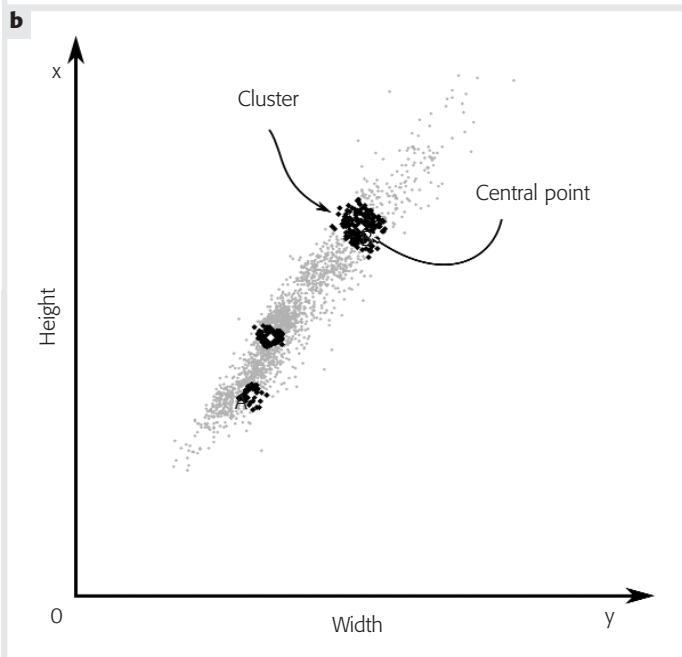
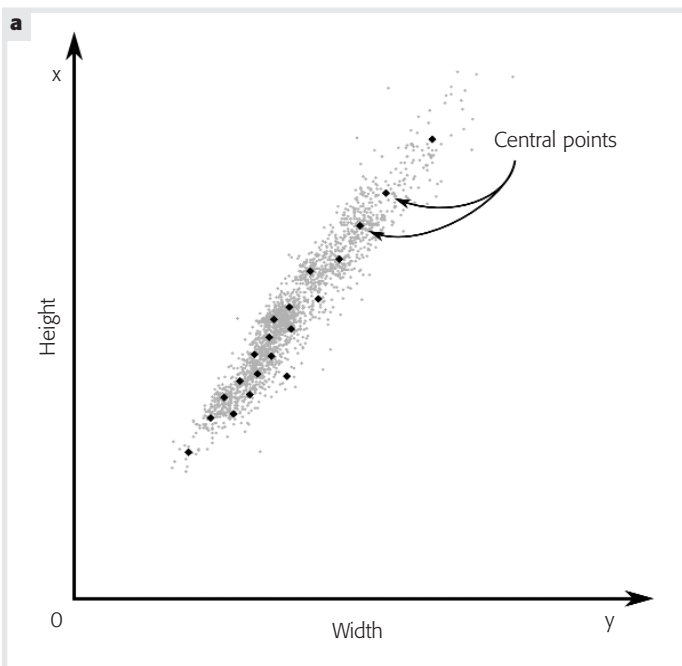
As mentioned in the description of the bin packing problem, the fact that all boxes in a stack have the same width means that the only dimension which concerns the packing is the height of the box. By employing the First Fit Decreasing algorithm we first sort the boxes according to height from the tallest to the shortest. Then we continue by adding the tallest boxes not yet stacked, to the first available stack that they fit in. We allow vertical margins between boxes as shown in Tab 2. We also check that the next shelf is aligned to a set of holes on the rack. We do that by

Tab 4 Sample data from the St. Catherine's manuscript collection.

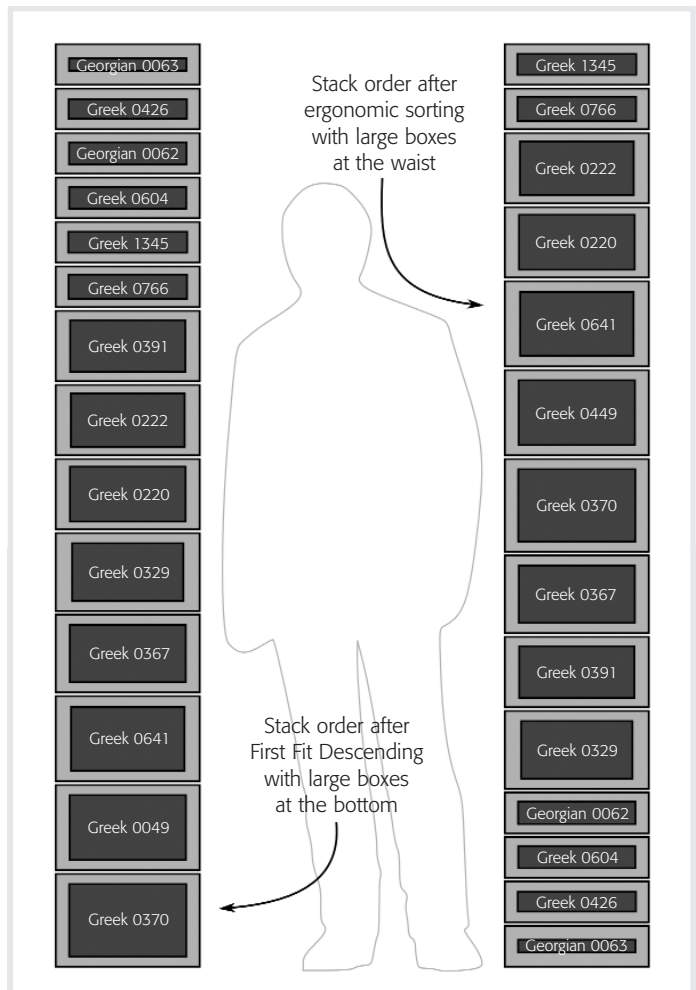
Shelfmark	Height	Width	Thickness
Arabica 0001	244	178	70
Arabica 0002	275	179	76
Arabica 0004	273	178	80
Arabica 0005	234	160	67
Arabica 0006	219	155	43
Arabica 0009	276	192	102
Arabica 0010	278	205	84
Arabica 0011	180	131	61
Arabica 0012	219	165	34
Arabica 0013	270	184	93

checking that the distance of every new box from the bottom of the stack is a multiple of the hole increment (e.g. for 20 mm increments, these are 20, 40, 60, 80 etc.). If this is not the case we increase the height of the previous box and maintain a constant margin for aesthetic reasons, therefore the gaps between boxes are always the same.

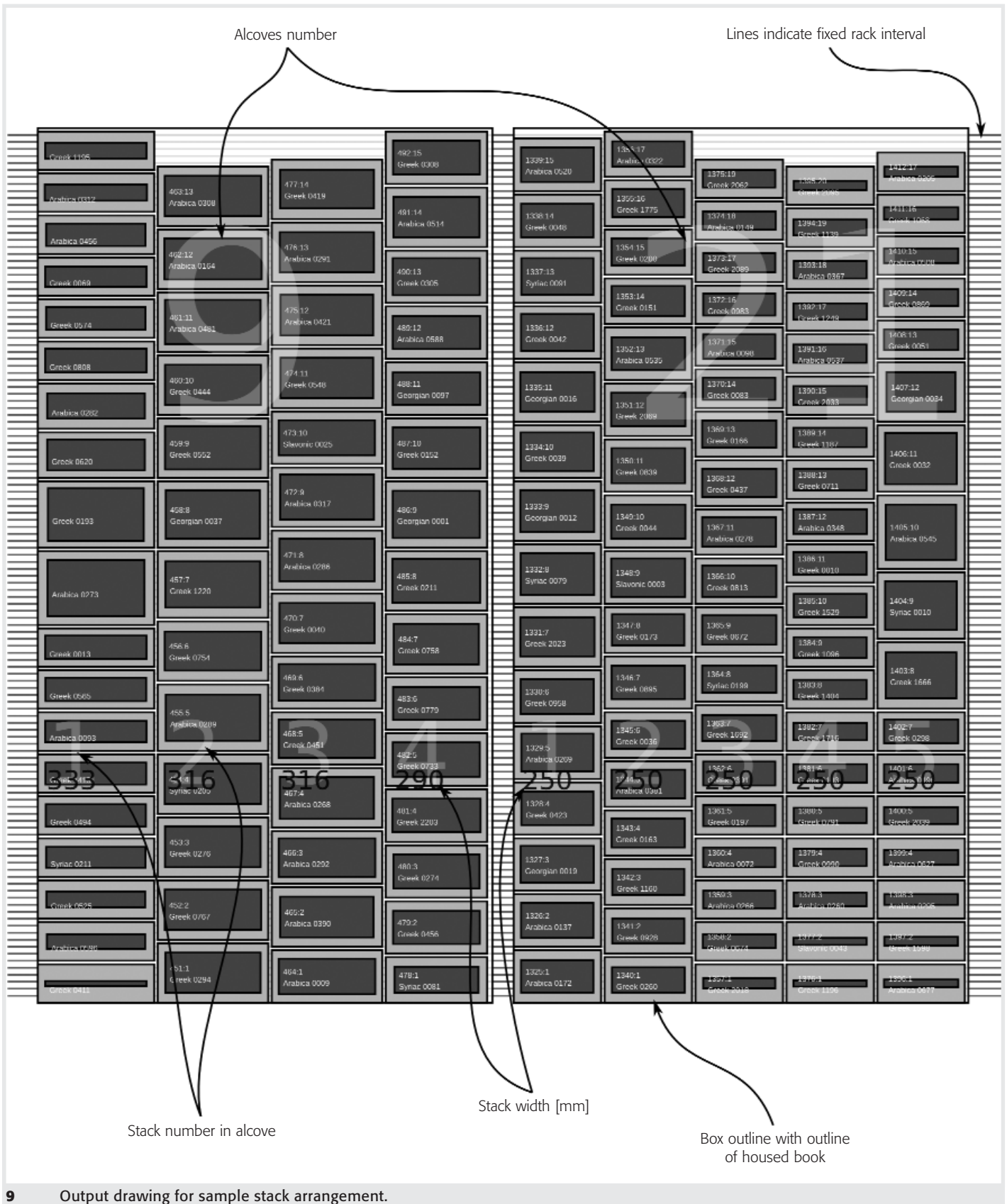
We only begin using a new stack when none of the existing ones can fit the next box. After all the boxes have been allocated to a stack, we repeat this process by fitting stacks to the available alcoves. We first sort the stacks according to width from the



7 Book width and height data after clustering: book dimensions data with central points after clustering (a); book dimensions data with selected clusters and corresponding central points (b).



8 Example stack before and after ergonomic sorting.



9 Output drawing for sample stack arrangement.

widest to the narrowest. Then we continue by adding the widest stack not yet in an alcove to the first available alcove that they fit in, until all stacks have been allocated to an alcove. We make sure that margins between stacks are taken into account and we centre the stacks in the alcove when there is unoccupied space again for aesthetic reasons.

As mentioned in the description of the stacks, because of the weight of large books, it may be ergonomically difficult to remove a large box from a stack when it is placed too low. The First Fit Decreasing solution typically places the largest boxes at the bottom of the stack, thus making the arrangement of the boxes less ergonomic. As a final step in our solution we look at

Tab 5 Sample data of calculated box sizes.

Shelfmark	Box height	Box width	Box thickness
Greek 0501	130	461	450
Greek 0356	190	461	450
Greek 0205	210	461	450
Greek 1613	190	461	450
Greek 0365	190	461	500
Greek 0499	210	461	450
Greek 1117	210	461	500
Greek 0379	210	461	500
Greek 0360	210	461	450
Greek 0375	210	461	450

the boxes in a single stack independently from the rest of the stacks. We re-arrange these boxes by filling the stack starting from the middle, so that the largest ones are placed at waist height. The overall height of the stack remains the same, so that this process does not, therefore, affect the bin packing result, but it makes the arrangement in each column more ergonomic (Fig 8).

Java Application

The calculation of box sizes and their arrangement on stacks can be done in a series of separate steps following the techniques described above. This means that a range of software tools is required when the parameters change. For example a different number of box sizes would require running WEKA again and exporting the resulting data manually for further processing. This makes the process slow and prone to errors. In order to automate and simplify the process, a new software application was developed which loads the book and alcove sizes from comma separated files (.csv)—as shown in Tab 4—and returns a graph in scalable vector graphics format (.svg [5]) with the box stacks alongside a table with box dimensions (Fig 9 and Tab 5). It does that by taking into account parameters which can be set on demand, thus easily including or excluding various steps in the process. The application was built using the Java programming language [6] on a NetBeans [7] environment. Java is a

popular open-source and well-supported language which allows resulting applications to run on multiple platforms (including MS Windows, MacOS and Linux). NetBeans is again popular and open-source, and thus a reasonable choice for us.

The application developed makes use of the k-means clustering and the First Fit Decreasing solutions to produce the resulting arrangement of boxes. A screenshot of the user interface of the application is shown in Fig 10. The various controls on the application window define the parameters on which the process is run and are simple to set. The fields around the box diagram control the linings and margins of the boxes. The two lists on the top right of the application show the dimensions of the books and available library space. The controls at the bottom right indicate the parameters for the k-means algorithm.

Conclusions

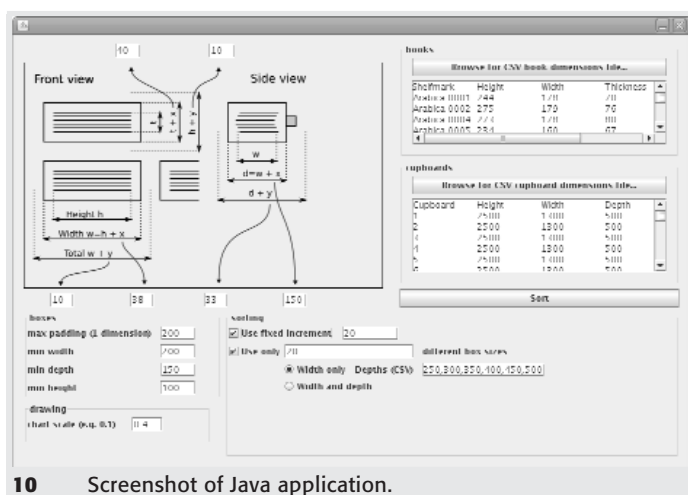
The problem of the calculation of box sizes and their spatial arrangement on stacks in libraries is not a simple one. An objective solution is impossible without making use of statistics on available book size data. The use of k-means and First Fit Decreasing solutions allow an objective solution to the problem and are tested methods in other fields of research. These techniques were applied to book size data from the St. Catherine's Monastery library. A new software application was used to simplify the process and to reduce the time needed to find a solution by removing any manual tasks. A good solution was produced on the St. Catherine's data and 20 different sizes of boxes have been established to accommodate the manuscripts. The proposed arrangement of the boxes on the stacks indicates that the boxed manuscripts will comfortably fit in the available library space as proposed in the architect's plans with only 27 of the 31 alcoves being occupied. This offers an important benefit to the whole design process since it confirms the suitability of the design for the specific collection. Similar tests could be run for any other collection where book dimensions are known and we intend to apply this methodology elsewhere and further report on our findings. We also intend to investigate other clustering and bin packing techniques and compare the results to identify an optimal solution for this and other collections.

Acknowledgements

The authors thank Prof John Harrison for reviewing the statistical methods proposed in this paper and Dr Nanke Schellmann for translating the abstract.

Endnotes

- [1] Ligatus is a Research Centre of the University of the Arts London <www.ligatus.org.uk>. The main objective of the centre is the study of historic bookbinding through the development of digital tools and resources with particular interest to conservation. During the IADA Symposium 'Out of Sight—Out of Mind' at Prague (27/28 May 2010), Nicholas Pickwoad held a contribution entitled 'Book boxes: A new design in stainless steel'.
- [2] Each book is placed inside a wrapper before being stored in the steel box. The book is not shown in the figure to simplify the drawing.
- [3] Mathematically this is defined as follows (e.g. Wikipedia 2011): Given a set of observations (x_1, x_2, \dots, x_n) , where each observa-

**10** Screenshot of Java application.

tion is a d -dimensional real vector, k -means clustering aims to partition the n observations into k sets ($k \leq n$) $S = \{S_1, S_2, \dots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS): $\arg \min \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$ where μ_i is the mean of points in S_i .

- [4] A good starting point for understanding the concept of NP-completeness and polynomial time is the book by Goldreich 2010.
- [5] <www.w3.org/TR/2011/WD-SVG11-20110512>.
- [6] <www.java.com>.
- [7] <<http://netbeans.org>>.

References

- Wikipedia (2011): k -means clustering <http://en.wikipedia.org/wiki/K-means_clustering>.
- Brock, Sebastian (1996): A venerable manuscript collection. In: Oriana Baddeley and Earleen Brunner (eds): *The Monastery of Saint Catherine*. London: Saint Catherine Foundation.
- Chen, Jie (2005): Comparison of clustering algorithms and its application to document clustering. Princeton, NJ: Princeton University.
- Coffman, E. G., et al (1996): Approximation algorithms for bin packing: A survey. In: Hochbaum, D. (ed): *Approximation algorithms for NP-hard problems*, Boston: PWS Publishing, pp 46-93.
- Gauch, Hugh G., and Whittaker, Robert H. (1981): Hierarchical classification of community data. In: *Journal of Ecology*, vol 69, no 2, pp 537-557 (Article-Type: research-article / Full publication date: July, 1981 / © British Ecological Society 1981).
- Gentleman, Robert, and Ihaka, Ross (2010): GNU-R <www.r-project.org/index.html>.
- Goldreich, Oded (2010): *P, NP, and NP-Completeness: The Basics of Computational Complexity*. Cambridge: Cambridge University Press.
- Hadgraft, Nicholas (1994): Storing and boxing the parker library manuscripts. In: Nicholas Hadgraft and Katherine Swift (eds.), *Conservation and Preservation in Small Libraries*. Cambridge: Parker Library Publications, pp 132-141.
- Hall, Mark, et al (2009): The WEKA data mining software: An update. In: *SIGKDD Explorations*, vol 11, no 1.
- Jain, A. K. (2010): Data clustering: 50 years beyond k -means. In: *Pattern Recognition Letters*, vol 31, no 8, pp 651-666.
- Jain, A. K., et al (1999): Data clustering: a review. In: *ACM computing surveys (CSUR)*, vol 31, no 3, pp 264-323.
- Korf, Richard (2003): An improved algorithm for optimal bin packing. In: 15th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, 9 to 15 august 2003, vol 18, pp 1252-1258.
- MacQueen, James B. (1967): Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol 1, p 14.
- De Maesschalck, R., et al (2000): The mahalanobis distance. In: *Chemometrics and Intelligent Laboratory Systems*, vol 50, no 1, pp 1-18.
- Manaphes, Konstantinos (1990): *Sinai: treasures of the monastery of Saint Catherine*. Athens: Ekdotike Athenon.
- Pickwoad, Nicholas, and Honey, Andrew: Book boxes: A new design in stainless steel. In preparation.
- Pickwoad, Nicholas (2004): The condition survey of the manuscripts in the monastery of saint catherine on mount sinai. In: *The Paper Conservator*, vol 28, no 1, p 33.

Sander, Jörg, et al (2003): Automatic extraction of clusters from hierarchical clustering representations. In: Whang, Kyu-Young, et al (eds): *Advances in Knowledge Discovery and Data Mining, Proceedings of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD-03*, Seoul, 2003, April 30 - May 2, Heidelberg, Berlin: Springer Verlag (Lecture Notes on Artificial Intelligence), pp 75-87.

Skiena, Steve S. (1998): *The algorithm design manual*. New York: Springer.

Townsend, Gregg (2010): Bin packing <www.cs.arizona.edu/icon/oddsends/bpack/bpack.htm>.

Ubeyli, Elif Derya, and Dogdu, Erdogan (2010): Automatic detection of Erythematous-Squamous diseases using k -Means clustering. In: *Journal of medical systems*, vol 34, no 2, pp 179-184.

Velios, Athanasios, and Pickwoad, Nicholas (2005): The database of the St. Catherine library conservation project in Sinai, Egypt. In: *IS&T Archiving 2005 Conference*, April 26-29. Washington, DC, USA: IS&T.

Suppliers

Conservation by Design Ltd., Timecare Works, 5 Singer Way, Kempston, Bedford MK42 7AW. United Kingdom, Tel +44-1234-846300, Fax +44-1234-852334, www.conservation-by-design.co.uk (steel box design).

Authors

Dr Athanasios Velios studied art conservation at the Technological Educational Institute of Athens. He completed his PhD at the Royal College of Arts on the automatic reconstruction of fragmented objects using 3D computer models. He has been active in the field of Computer Applications to Conservation for many years developing computer software for conservation-related applications. He is currently working on the semantic documentation of historic bookbindings.

Dr Athanasios Velios, Ligatus Research Centre, CCW Graduate School, University of the Arts, London, 16 John Islip Street, London SW1P 4JU, United Kingdom, Tel +44-207-5146432, a.velios@gmail.com

Professor Nicholas Pickwoad trained in bookbinding and book conservation with Roger Powell, ran his own workshop from 1977 to 1989, and has been Adviser on book conservation to the National Trust from 1978. He was Chief Conservator in the Harvard University Library from 1992 to 1995 and is now project leader of the St Catherine's Monastery Library Project based at the University of the Arts London and is director of the Ligatus Research Centre.

Professor Nicholas Pickwoad, Ligatus Research Centre, CCW Graduate School, University of the Arts, London, 16 John Islip Street, London SW1P 4JU, United Kingdom, Tel +44-207-5146432, npickwoad@paston.co.uk

Andrew Honey is a book conservator at the Bodleian Library, University of Oxford. He graduated from Camberwell College of Arts in 1994 with a BA (Hons) in Paper Conservation and studied the conservation of rare books and manuscripts at West Dean College under Christopher Clarkson from 1995 to 1997. He has been a visiting research fellow at the Ligatus Research Unit, University of the Arts London since 2005 working for the Saint Catherine's Monastery Library Project.

Andrew Honey, Bodleian Library, Broad Street, Oxford OX1 3BG, United Kingdom, Tel +44-186-5277080, andrew.honey@bodleian.ox.ac.uk