Преглед НЦД 9 (2006), 1–11

Eugenia Stoimenova,
Plamen Mateev,
Milena Dobreva
(Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences)

OUTLIER DETECTION AS A METHOD FOR KNOWLEDGE EXTRACTION FROM DIGITAL RESOURCES

Abstract. Mass digitization leads to the gathering of large amounts of data and metadata in electronic form. Commonly, they are used for representation and data harvesting.

In information retrieval we have the cases of records, which differ much from the main part of the data. They seem to be quite unusual than one would expect from the rest of the records and from the "knowledge" about the underlying process, which generates the information items. Such records are usually called “outliers”. This information can lead to substantial improvements in the model. It can also lead to discoveries, which are valuable themselves.

The basic aim of this study is to demonstrate what knowledge could be extracted studying the outliers in a collection of Bulgarian mediaeval manuscripts metadata. The distribution of document size is investigated using statistical techniques. Several outliers were marked as misprints, some other were pointed as documents with non standard intention. The distribution of extent data showed a structure that might be explained by the paper folding preferences. An appropriate technique for distribution was utilized and the manuscripts were presented according to their chronological distribution.

Key words: cultural heritage, knowledge extraction, outlier detection, data verification.

1. Introduction

Statistical study of data gives a general impression on certain phenomena. We postulate a model analyzing the data features. The goal is to use it for prediction of the consequences of our actions and manipulation of our environment. In some sense the postulated model reflects what we already know or we think we know.

One characteristic feature of the digital preservation of and access to cultural heritage is the collection of voluminous data in electronic form which processing is still not a trivial task. These data most often are encoded within the metadata accompanying digital images and full texts. For example, such records are available in the cases of manuscript and archival descriptions, museum collections, etc. Currently, these data are used in visualization and in identifying records which answer specific criteria. The processing of these data as a data collection could lead to discovering new facts about the cultural and scientific heritage, for example its regional and chronological distribution, items which do not follow the general trends, etc. As an example of innovative approach applied to a collection of manuscript data we could mention the development of proper intelligent agents for search and processing purposes which are able to retrieve and filter data (documents and images) by their semantic properties [4].

In our work we present the application of outlier detection methods in the studies of data on cultural and scientific heritage resources. In Section 2 the statistical notion of outlier and methods of their detection are described. The data set of interest is presented in section 3. We present an empirical study of some useful variables on a sample of KT-DigiCULT-Bg Resources data sets in section 4 and 5. Applicability of selected methods for univariate and multivariate outlier detection ([2] and [7]) is tested. Comparative analysis and some unexpected facts are reported.

2. Outliers

2.1. Why do outliers appear? In many experimental tasks, in information retrieval, for instance, we have the situation of records which differ much from the main part of the data. They seem to be surprisingly different, higher or lower, than one would expect from the rest of the records and from the "knowledge" about the underlying process or semantics, which generate the information items. Such records that do not fit the overall trend are usually called “outliers” although no formal definition exists.

For instance, in the sequence 0, 1, -1, 103, 3, -2, the number 103 is an outlier. An outlier may indicate that there was an error in the process that produced the data, or it may show there is a real abnormality in the system that we are studying.

An "outlier" is a statistical term. It refers to an observation, record in the database that lies an abnormal distance from other values in a sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal in an assumed context. Before abnormal records can be singled out, it is necessary to characterize normal records.

Undoubtedly, it depends on the assumed model if an extreme record is considered as surprisingly different, i.e. if it arises from some other source than the remaining data. The main goal of any statistical analysis is a study of the norm or typical characteristics of a system. Presents of outliers render any standard statistical analysis difficult. Outliers might give bias impression about the system and could lead to wrong decision, for instance.

2.2. What should be done with outliers? When we encounter an outlier, we may be tempted to remove it from the analysis. In our consideration the data would be records of document description. Before deciding to remove it, we should ask these questions:

- Was the value entered into the computer correctly? If there was an error in the data entry, it should be fixed.
- Were there any problems with collecting data in that record? For example, if we noted that one document description looked false, we have justification to exclude its record without needing to perform any other action.
- Is the outlier caused by document diversity? If each record comes from a different chronological time, the outlier may be a correct value. It is an outlier not because of recording a mistake, but rather because documents from this time may be different from the others. This may be the most exciting finding in our data!

If the answers to those three questions are negative, we have to decide what to do with the outlier. There are two possibilities. One possibility is that the outlier appeared due to chance. In this case, we should keep the value for our analysis. The value came from the same population as the other values, so it should be included. The other possibility is that the outlier appeared due to a mistake – bad recording, forgery, imitation, etc. Since the presence of an erroneous value will cause invalid results, it should be removed from the analysis. In other words, the value comes from a different population than the rest and is misleading.

The problem, of course, is that we can never be sure which of these possibilities is the correct one. Clearly, no mathematical calculation will tell us for sure whether the outlier came from the same or different population. However, statistics can answer questions like such as: If the values really were all sampled from a given distribution, what is the chance that we would find one value as far from the others as we observed? If this probability is small, then we may conclude that the outlier is likely to be an erroneous value, and we have justification to exclude it from our analysis.

2.3. Methods for outlier detection. Many methods have been proposed for univariate outlier detection. They are based on robust estimation of location and spread, or on quantiles of the data. A major disadvantage of these methods is that the decision rules are independent from the sample size. Moreover, by definition of most rules (e.g. based on a distance from the mean) outliers are identified even for "clean" data, or at least no distinction is made between outliers and extremes of a distribution.

Graphical techniques like histograms, box-whisker plots, as well as X-Y scatter plots, are the common tools for finding interesting subjective cases (Figures 1, 2, 3) or existence of unexpected structure (Figures 4).

All detection methods first quantify how far the outlier lies from the other values. This can be the difference between the outlier and the mean of all points, the difference between the outlier and the mean of the remaining values, or the difference between the outlier and the next closest value. Next, standardize this value by dividing by some measure of scatter, such as the standard deviation (SD) of all values, the SD of the remaining values, or the range of the data. Finally, compute a corresponding probability (*p-value*) answering this question: If all the values were really sampled from a Gaussian population, what is the chance of randomly obtaining an outlier so far from the other values? If the p-value is small, we conclude that the deviation of the outlier from the other values is statistically significant.

The basis for multivariate outlier detection is the Mahalanobis distance. The standard method for multivariate outlier detection is robust estimation of the parameters in the Mahalanobis distance and the comparison with a critical value of the Chi-square distribution [7]. However, values larger than this critical value are not necessarily outliers, they could still belong to the data distribution.

3. Data sets

At the present moment, the work on KT-DigiCult-Bg project [3] foresees collection of metadata on mediaeval manuscripts, archival records (jointly with the General Department of Archives) and mathematical publications of Bulgarian authors.

Our study is based on a sample of metadata on Bulgarian manuscripts in Bulgarian repositories. Currently we have 806 descriptions in XML format conformant to the TEI P5 DTD, most of them are from the catalogue [1]. Details on this effort are given in [5]. The number of descriptions will increase with project evolution and will lead to large data sets of different types.

All these resources contain large data sets of different nature. Some of them contain structured information and metadata, as well chaotic collections of different size text segments. Using more or less intelligent retrieval systems one can extract relatively small subset of items with homogeneous structure. The extracted elements may be presented as vector of features. The text segments usually are presented via indicators of some key words with specified frequencies (of words, collocations, letters, bigramms, trigramms etc.), i.e. as vectors of features too.

The data set is a multivariate comprised different features of the manuscripts.

The documents are grouped according to their preservation status and folio material. The distribution of the documents in these groups is described bellow.

Manuscript status defines the following categories of documents:

- Unitary document (complete entity which exists as a single unit). It numbers 47 documents.
- Composite document (multiple units of different origin). There are 94 documents.
- Fragmentary (a few sheets, a small part of a sheet, or a manuscript). It numbers 81 documents.
- Defective document (minority of the leaves are missing). This is the largest part of documents, contains 343 manuscripts.
- 242 or 30% of all items have unmentioned document status.

The presence of defects is observed in 43% of all documents. Another significant part of documents are those with quite normal status which is not noted and signed as "unknown". In most considerations the documents from different classes do not show different features. So the documents are combined in a single sample if it is not specially mentioned.

Folio Material defines three categories, namely **paper**, **parchment** and mixed **paper and parchment**. The distribution of documents according folio material is as follows: most of the manuscripts, 725, are written on paper; only 78 or nearly one of ten (9.7%) is on parchment; a quite small part of documents are written on a mix of two materials (its number is 4).

Other manuscripts' features are characterized by **quantities**. Our experiments are based mainly on the physical characteristics of the documents. The corresponding quantitative variables are: **extent** – the number of folios in the item; **height** and **width –** horizontal and vertical sizes of the folio in [mm].

Date of origin of manuscripts is a key feature. In most cases the documents are not dated exactly. The verbal descriptions are not unified and contains various forms of dating, ranging from a specific year to a span of 2 centuries. The verbal description looks like: "at the {end/middle/beginning} of the YY century" or "at the {first/second} half of the YY century".

In order to process these data we transform verbal descriptions to intervals covering the period when the manuscript is written. The new variables are "Notbefore" and "Notafter".- time period in which a manuscript could be likely written. This corresponds to accepted way of origin date description.

4. Experiments and Results

4.1. Finding unusual extent of documents. In many situations outliers are easily handled and the manner of dealing with them is obvious. Such is the situation when human errors lead to incorrect recording of data. The univariate variable Extent comprises of the number of folios in the document. In a study of this variable few obvious misrecorded values were corrected. Further we study Extent for unusual extreme values.

The values have large amount of variability. Examination of the overall shape of the plotted data includes study of symmetry and departures from normal assumptions. The distribution "looks" natural and it is not expected to be normal. We observe that the number of folios have asymmetric distribution with right skewness. The shape is illustrated by a **histogram** plot.

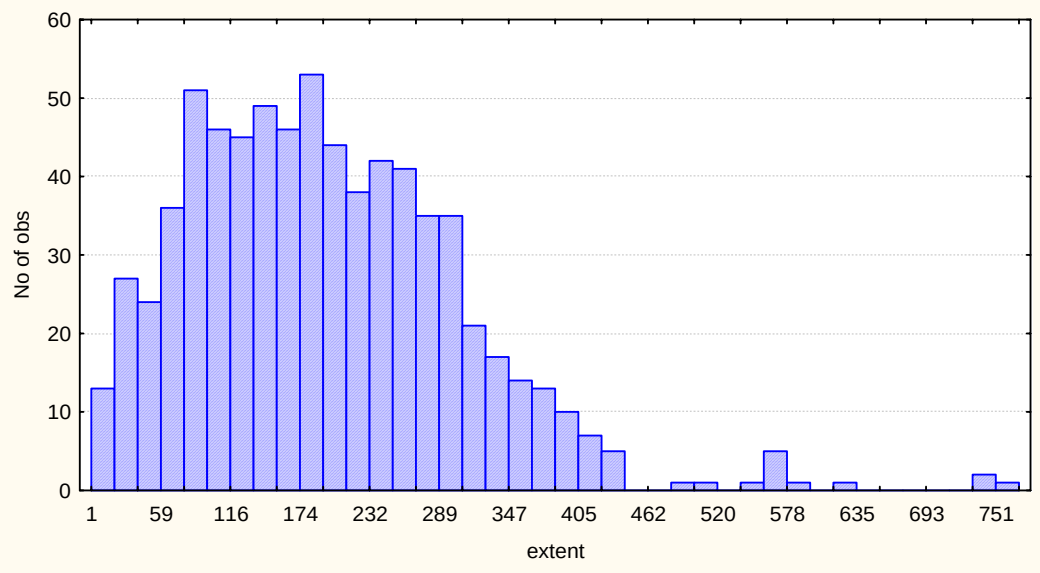

Figure 1: Histogram of the ***Extent*** presents the distribution of the number of folios in manuscripts described in the data base.

It seem that values larger than 450 approach to unusual large extent of a manuscript. The question is "Is the outlier caused by document diversity, or this is a recording mistake?"
The box plot is a useful graphical display for describing the behavior of the data in the middle as well as at the tails of the distributions. The box plot uses the median and the lower and upper quartiles (defined as the 25th and 75th percentiles) (Figure 2):

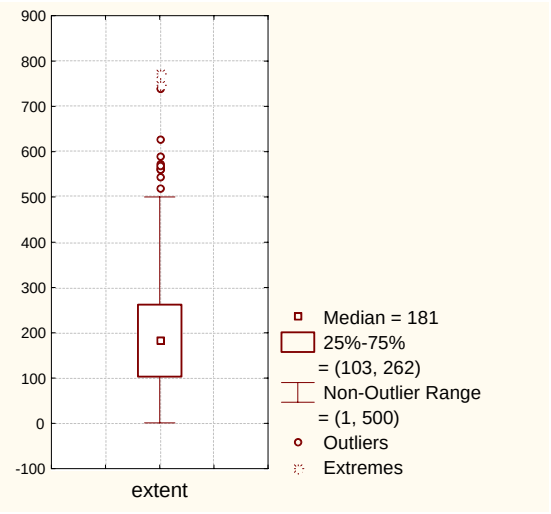

Figure 2: Box-plot presents distribution using median, quartiles, minimal and maximal values.

Examination of the data for unusual observations that are far removed from the mass of data is done by statistical tests like Tukey's. Tukey's rule is based on the interquartile range – the difference (*Q2 - Q1*), where *Q1* and *Q2* denote the lower and the upper quartiles [2]. A data point is deemed to be an outlier if the following conditions hold:

data point value > *Q2 +1.5*(Q2 – Q1)* or
data point value < *Q1 -1.5*(Q2 – Q1)*

A data point is deemed to be an extreme value if its value departures from the two quartiles by three interquartile ranges (i.e. the factor 1.5 is replaced by 3 in the above two inequalities).
The Tukey's rule defines the outlier upper bound of ***Extent*** as 440. The documents in our database that are determined as possible outliers are 13 and part of their description is given in the Appendix. Seven out of them are kept in Rila monastery library. Nearly the half of these manuscripts is composites. Processing of outliers (or spurious values of any sort) in such cases is not a matter of statistical analysis, but of native wit.

Remark on the assumption for normal distribution: If a normal distribution is assumed then a single value may be considered as an outlier if it falls outside the range of 2.5 standard deviations around the average:

[average ± 2.5*stdev]

This gives about 99% confidence interval. The popular Grubbs' test is more precise for small samples and it is based on Student's *t*-distribution [2].

4.2. Study of the manuscript format. Data set of consideration includes two physical characteristics of the documents – the dimensions **height** and **width**. The values of two variables have remarkable amount of variability and clustering. Separately study of dimensions discovers one outlier with 840 mm height (manuscripts ID 00755). On further enquiry, however, it was found to be perfectly reasonably: the manuscript is a scroll! It was excluded of further consideration.

There was a hypothesis that dimensions of documents depend on the support material (Parchment / Paper). The descriptive statistics minimum, lower quartile, median, upper quartile and maximum characterize the distribution of the dimensions of the folios:

Manuscripts:		**min**	**Q1**	**Med**	**Q2**	**max**
Paper	Width	64	138.5	160	202	240
N=721	Height	87	195	220	290	360
Parchment	Width	90	150	175	195	300
N=76	Height	140	210	245	270	360

The parchment documents are not so many and their characteristics look similar to paper ones. Actually, there is no statistically significant difference between the characteristics of the two categories documents. Thus, we continue with paper documents only.

The study of outliers has as much relevance and importance for multivariate data as it does for univariate samples. Figure 3 plots the height of the folios versus its width. There is a strong linear relation between the two sizes. The scatter plot reveals two multidimensional outliers. Namely:

ID00467 – (height = 305, width= 105)

(Народна библиотека "Св. Св. Кирил и Методий", София; Катасник (Поменик на дарители) на Черепишкия манастир);

ID00078 – (height= 165, width= 200)

(Народна библиотека "Св. Св. Кирил и Методий", София; Четвероевангелие; 14 век, средата)

The current finding points researchers to manuscripts with untypical sizes in the database.
For these particular manuscripts none of the component sizes is "surprising" in relation to its univariate distribution and yet theirs assemblage of measurements as multivariate observations seems "surprisingly far away" from the main group of data.

What would be the explanation of these facts? Are they mistakes in recording the data or there are more sophisticated reasons for them? Further explanation will be due to medievalists.

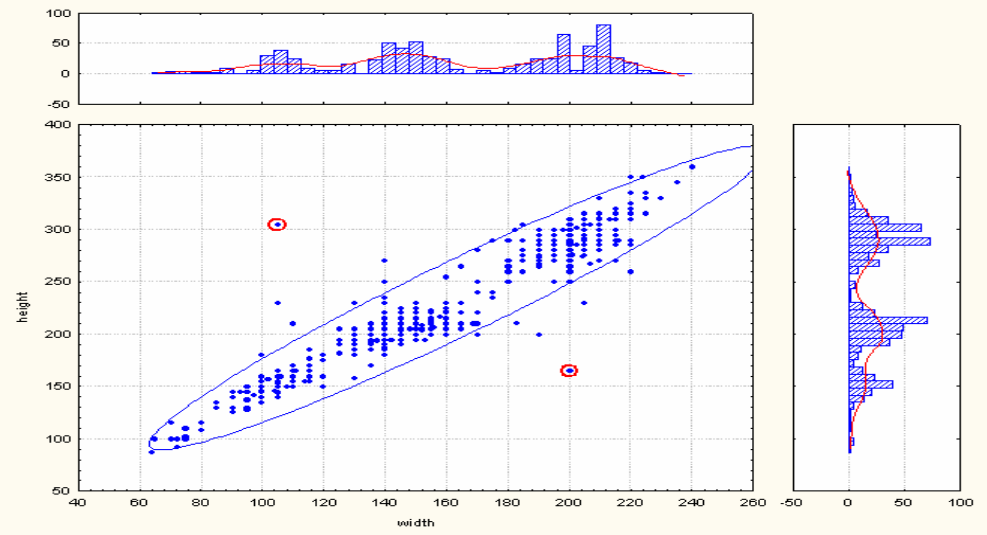

Figure 3: The X-Y (***width-height***) scatter plot of paper documents reveals two multidimensional outliers.

4.3. Clustering of the Paper documents. The examination of univariate samples has somewhat limited aims and utility. More often, and more usefully, we need to consider more structured situation. However, outliers are not of interest here, similar techniques are applied to discover some unknown structure in data. The histograms of width and height on Figure 4 reveal al least three remarkable clusters.

It seems (Fig.4) that the three "typical" book formats are describe by:

Small: height in [120, 175] and width in [90, 120],
Medial: height in [175, 245] and width in [120, 175],
Large: height in [245, 345] and width in [175, 245].

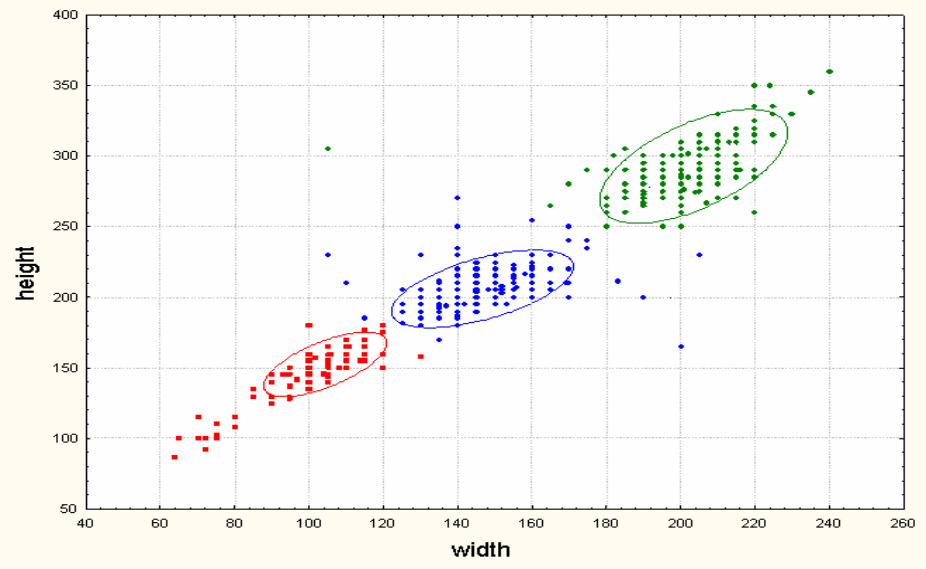

Figure 4: The three clusters on the X-Y (***width / height***) scatter plot of paper documents.

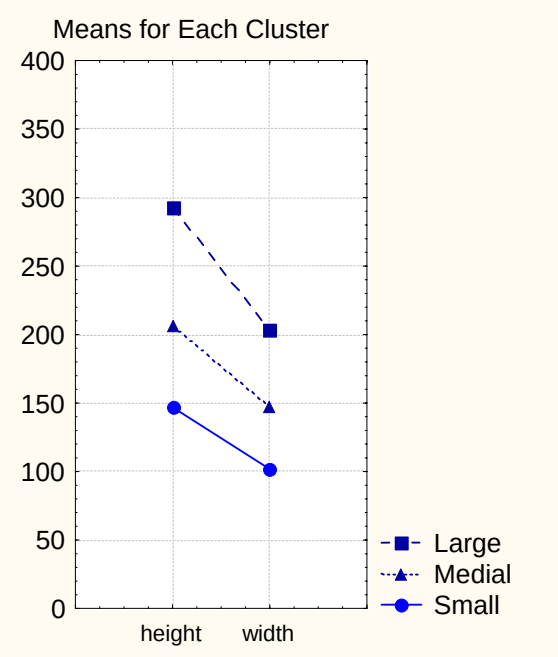

Figure 5. Height and width averages of the three clusters.

This hypothesis was tested using the "k-mean" clustering procedure. Three clusters were determined of similar sizes (133 small, 323 large and 265 medial). The diagram of averages of dimension variables (Fig. 5) for three clusters show a relation between average document dimensions.

The height average of medial size books equals to width average of large books and the width average is equal to height average of small books. The small ones have dimensions twice smaller than those of large books.

A natural explanation of this clustering is the way of folding books. According an explanation of E. Krushelnitzkaya (in private communication) the books might be produced in "standard" formats. The folio size of manufactures determines the book "a page" format. The size of the sheet should be that if you fold it once you have a book in "a half page" format, and if you fold it twice, you have a book in "a quarto" size format (Fig. 6).

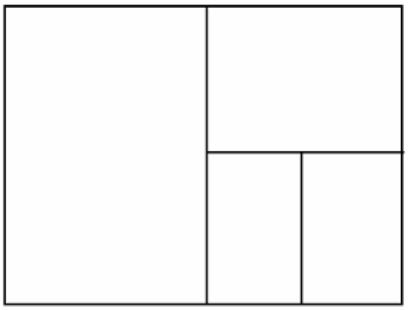

Figure 6. The manner of folding sheets in "a page" format, "a half page","a quarto" size format

This is possible if the ratio of two dimensions of the sheet is the same as ratio of side and diagonal of the square or the ratio have to be equal to squared root of two. Thus this ratio is saved the same after folding The tradition is retaining in modern paper standard formats (…, A3, A4, A5,…), which ensure convenient variety with minimal loss of paper.

The question what standard sizes of folios were produced or used in the Slavonic lands is still interesting for systematic investigation of medievalists. Our hypothesis is that dimensions of the folios and frames had been 41 and 29 cm nearly, enough for two large pages.

5. Which books survived within different ages?

The goal of this section is to present a method of describing imprecise quantitative temporal data. As we mentioned above, the dating of manuscripts is not exact and is presented by "Notbefore" and "Notafter" – time period in which a manuscript could be likely written. An appropriate representation of data could help in discovering data structure and outliers as well. In the example below we illustrate the distribution structure of origin date.

5.1. Distribution of preserved manuscripts over time. Usually data distribution is analyzed graphically. The simple histogram plot for the studied time period is not directly applicable to such type of data. The modified histogram of transformed origDates used the "Notbefore" and "Notafter" instead the exact years (Fig.7). The collected data are a sequence of independent intervals. We suppose that a document is could be produced at any time of its interval and the total mass in the interval is 1.

The distribution of transformed data is even better presented by modified smoothed histogram. Further, let [*a,b*] be a bin for histogram construction. Analogously to counting, we define a mass function *m(a,b)* that restricts the probability mass of the recorded intervals over the bin [*a,b*]. For each interval *A*, the restriction is the fraction of *A* overlapping [*a,b*]. Thus *m(a,b)* is the restricted mass of all elements of the sample into this interval.

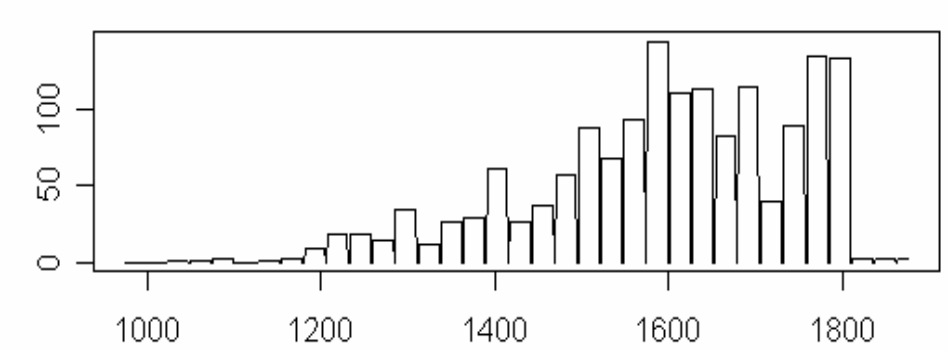

Figure 7. Histogram of OrigDate.

Therefore an empirical density function (smoothed histogram) in the interval [*a,b*] is given by

$$f(x) = m(a,b)/n, \quad \text{for } a < x < b.$$

Since it is reasonably to assume that the mass changes smoothly over the studied period, the mass falling in one particular time interval provides information about the probability of falling in its neighbors. Therefore, smoothing makes sense since we assume that the distribution is continuous. The improvement using smoothing is most evident when the distribution is sparse in the sense that mass falling in each histogram bin is small. More details about the method of smoothing will appear in [6].

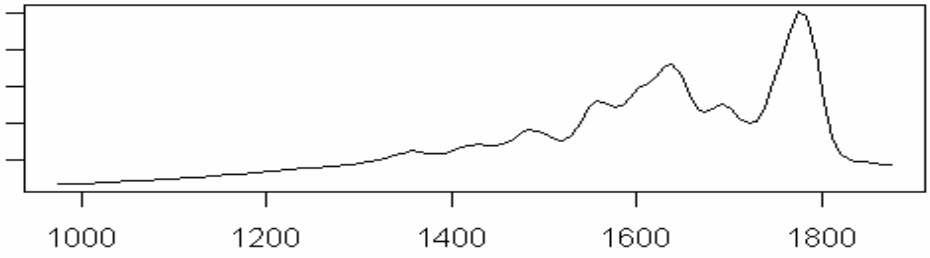

Figure 8. Smoothed histogram of OrigDate.

The last diagram (e.g. Fig 8) shows a tendency of increase of the number of described manuscript with the time. There are two exceptions in the trend. The first one, at beginning of 16-th century, is smaller. The second one is more significant and more prolonged at the end of 17-th and beginning of 18-th century. Such effects can not be denoted on unsmoothed histogram.

5.2. Distribution of clusters over time. The distribution of manuscripts over time axes reveals another feature of the manuscript collection. Smoothed histograms on Figure 9 are constructed for "Notbefore" and "Notafter" dates for the three clusters obtained in 4.3 separately.

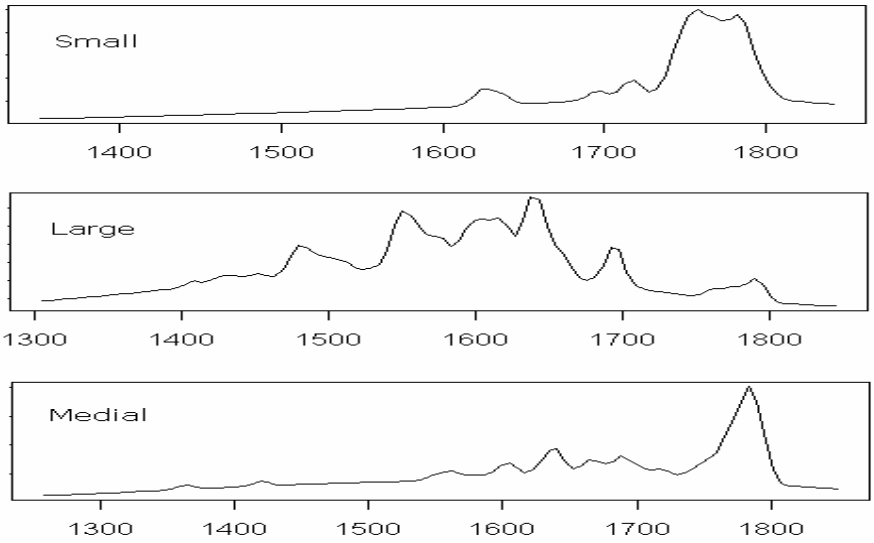

Figure 9. Smoothed histogram of **OrigDate** for the three clusters.

Small size books are mainly from the last centuries (1730 – 1800) while large size books are distributed in earlier times (most of them between 1450 – 1650). The medium size books are distributed more uniformly over the years (with a peak at the end of 17-th century). Most likely the massovization of book usage in the last centuries lead to the preference to small books, easy to carry during travels.
The question what standard sizes of folios were produced or used in the Slavonic lands is still interesting for systematic investigation of medievalists.

Conclusions

We have showed that outlier's detection technique is an important step in the management of manuscripts data base. The presence of outliers in manuscript data sets arises by two common reasons: errors in the data entry or extreme values due to diversity. Treating the data with outliers may give bias impression about the features of the total collection of manuscripts data. After outliers detection a decision about any of them should be taken. If an outlier appeared due to chance it should be kept for further analysis. If an outlier appeared due to a bad recording, it should be removed from the analysis. We have demonstrated that new knowledge could be extracted studying outliers in collection of Bulgarian mediaeval manuscripts metadata.

Acknowledgement. This research has been supported by a Marie Curie Fellowship of the European Community programme "Knowledge Transfer for Digitalization of Cultural and Scientific Heritage in Bulgaria" under contract number MTKD-CT-2004-509754.

References

[1] Икономова, А., Д. Караджова, Б. Христова. Български ръкописи от XI до XVIII век, запазени в България. Своден каталог, том I, НБКМ, София, 1982.
[2] J. D. Jobson, Applied Multivariate Analysis, Springer-Verlag, 1991.
[3] KT-DigiCult-Bg project, MTKD 509754.
[4] M. Nisheva-Pavlova, P.Pavlov, Tools for Intelligent Search in Collections of Digitized Manuscripts. In: "From Author to Reader", 9th ICCC International Conference on Electronic Publishing (eds. M.Dobreva, J. Engelen), Peeters Publishing, Leuven, 2005, 145–150.
[5] P. Pavlov, XeditMan: A XML Editor for Manuscript Descriptions and Its Implementation for Cataloguing of Bulgarian Manuscripts. NCD Review, 5 (2004), 49–58.
[6] E. Stoimenova, Empirical density estimation for interval censored data. 2005, (unpublished).
[7] D. Vandev, Detecting multidimensional outliers, Technical report, 2004 (unpublished).

jeni@math.bas.bg
pmat@math.bas.bg
dobreva@math.bas.bg

Appendix The following documents are possible outliers:

	recordID	msDescription_status	msIdentifier_repository	msIdentifier_altName
(770)	00793	compo	Читалище "Искра", Казанлък	Триод цветен
folios	00131	compo	Национален музей "Рилски манастир"	Панегирик (Рилски)
(738)	00634	compo	Народна библиотека "Св. Св. Кирил и Методий", София	Сборник (Янкулов)
(628)	00466	uni	Народна библиотека "Иван Вазов", Пловдив	Сборник
(590)	00641	def	Народна библиотека "Св. Св. Кирил и Методий", София	Сборник от слова и жития
(572)	00213	uni	Национален музей "Рилски манастир"	Сборник от жития
(569)	00353	unknown	Великотърновска митрополия, Велико Търново	Миней празничен и триод цветен
(565)	00130	compo	Национален музей "Рилски манастир"	Сборник "Андрианти" на Йоан Златоуст
(560)	00132	compo	Национален музей "Рилски манастир"	Панагирик
(559)	00195	def	Национален музей "Рилски манастир"	Шестоднев на Василий велики и тълкувания на Теофилакт Български
(542)	00281	unknown	Национален музей "Рилски манастир"	Патеричен сборник

(517)	00576	def	Църковен историко-археологически музей, София	Сборник богослужебен