

BAR-ILAN UNIVERSITY

Using Data Mining Techniques for Analyzing Pottery Databases

Zachi Zweig

Submitted in partial fulfillment of the requirements for the Master's degree in the Martin
(Szusz) Department of Land of Israel Studies & Archaeology, Bar-Ilan University.

This thesis was written under the
supervision of Prof. Aren Maeir

Abstract

Archaeological research has always been data driven. Now-a-days, due to the rapid development of computing technology, the use of archaeological databases is starting to take a major role in archaeological studies. Although the amount of archaeological data has increased significantly, efficient methods for analyzing this data are still absent. The Data Mining process may resolve this problem by helping us transform the data into significant information. This thesis deals with testing the possibility of implementing Data Mining techniques on archaeological databases, while focusing on pottery data.

Data Mining is the science of extracting useful information from large data sets. It was necessary to develop various unique methods for creating and preparing the data for this kind of process and analysis. Wide data tables with many variables were created. The methods by which this was done, and their implementation, added another aspect to this research. The typological issue was also dealt with from various different angles, while trying to find methods whereby ‘real’ types would be revealed and identified. The purpose behind the various methods used, was not to accept nor to reject any existing hypotheses, but rather to help find certain patterns and relationships within the data, which could prove worthy of further investigation as to their archaeological meaning.

The research was done on two case studies. The primary one was based on 280 whole vessels that were found at the site of Tell es-Safi (the biblical Philistine town of Gath). In this case study, different methods for the detailed recording of the pottery were also examined. This data had sampling problems because of its relatively small size, and lack of spatial and chronological variability. The second case study focused on processing and analyzing data that had already been published. Here a large data table of about 8000 pottery shards that were found at the site of Tel Batash (the biblical town of Timnah), was used.

In the **first chapter** the different Data Mining processes as they are used in the SAS Enterprise Miner 5.1 software, are described. This software makes use of the SEMMA (Sampling, Exploring, Modifying, Modeling and Assessing) process, and uses a process flow diagram for setting the flow of the various nodes. One of the primary nodes is the Metadata tool, which gives information about the variables of the data. Each variable is

defined by its Role (Input, Target, ID, etc.) and Level (Interval, Nominal, Ordinal and Binary). Another important node that is commonly used in this research is the Decision Trees tool. This method creates an empirical tree that represents segmentation of the data, by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in an hierarchy of segments within segments. These rules can be verbally described, unlike the other kinds of models which use a 'black box' approach.

In addition, other nodes were used, such as: Impute – for filling in missing values; SAS Code – for incorporating code into the process flow and thereby extending the functionality of the process; Transform – for creating new variables from existing ones; Cluster Analysis – for clustering similar vessels and discovering types. Furthermore, the Data Mining Neural Networks node, which is based on the Neural Networks theory, was implemented to produce fuzzy variables, by creating prediction models for selected attributes.

In addition to the Data Mining tools, a few simple Exploratory Data Analysis methods that were used for analyzing and displaying the Data Mining results, are also presented. These methods included: Contingency Tables (Chi-Square tests), Box-n-Whisker plots and Histograms.

The **second chapter** covers the topic of the use of computerized databases in archaeology. This issue became one of the major research challenges with which I had to deal. In this chapter the accelerated development of the use of databases in archaeology in the last two decades is surveyed. Nevertheless, there is still an absence of formal protocols for recording archaeological data. The problems of objective recording, and that of using categorical-typological attributes, are introduced. Two questions arise: How detailed should this recording be? and: Which are the important traits that should be recorded? Emphasis was placed on the substantial potential of the Relational Database which enables enrichment of the data with contextual information. Furthermore, some new methods for recording the morphological characteristics of a vessel are introduced. These methods provide objective ways to record these kinds of attributes. The Karasik and Smilansky method is also included because of its use in the current research. With this method the shape of the vessel profile is represented as a mathematical curve

function. Using this function, vessel shapes can easily be compared, and various morphological attributes can be extracted.

In the **third chapter** the different processes that were used for creating the data table of the Tell es-Safi vessels are described. To start with, all the drawings of the vessels were scanned, and the images were processed in a way that enabled importation into MATLAB software. The Karasik and Smilansky programs were then run on this data. Following that, using Microsoft ACCESS, a special form was created for entering all the data relating to measurements that were directly and manually taken from the vessels. In the first phase the vessel drawing was incorporated within the form, and attributes that were seen in the drawing were recorded. Some of these attributes were defined typologically, such as the form of the rim or of the base. However, other attributes were easier to measure, such as rim direction or number of handles. A secondary form was used to record decoration attributes. In the next phase, attributes measured from the actual vessels were recorded. Such attributes were: hardness, color, slip preservation, erosion level, etc. In addition, some basic morphological attributes were measured, since some of the whole vessels had not been drawn. The next phase consisted of recording traits measured by macroscopic analysis of the clay fabric. A small shard sample was taken from most of the vessels for this analysis. By examining a freshly broken section with a magnifying glass, various attributes were measured, such as: types of inclusions, their frequency, size, shape, color and luster, or voids types, and their frequency, shape and size. Additional attributes were measured by other means, such as, differences in color between core and margins, core hardness, reaction to hydrochloric acid, reaction to a magnet, specific weight, etc.

In these forms, there were instances where special text fields had to be created. These fields were necessary for describing complicated attributes that could not be described using lists of categories for nominal variables, such as: scraping marks, inclusions, changes in oxidization levels along the section, etc. Consequently a unique language code for describing each trait was created. These values were later extracted into simple variables using special programs which had to be specially written for this purpose. .

The final phase of the manual recording included contextual information. The temporary stratum number for each locus that was used in the new vessels table had to be

entered. For better spatial analysis, different architectural units in the architectural plan had to be defined and assigned to the loci in use in the table.

The next stage involved computing many morphological attributes. These computations made use of the Karasik and Smilansky method. Using the MATLAB software, various programs were written for the purpose of automatically defining the borders of different parts, such as: lip, rim, shoulder and base. In addition, the profile was automatically subdivided into elemental sections using different methods. Two of them were based on curve peaks, one was based on inflection points, and another on curvature changes in relation to the horizontal axis. Many attributes were calculated for each section, such as: average curvature, average thickness, curvature skew, curvature kurtosis, section relative length, etc. This resulted in the creation of a total of 946 morphological variables.

The final stage of preparing the data was done by programs written with the SAS Base language. The different data tables were merged into one table. The contextual data was also merged into this table, and in addition to the regular locus card fields, it included information on various artifacts that were found in the vessel's locus. This resulted in a data table which included a wealth of contextual information. For example, for each vessel there was information about the feature in which it was found, and other artifacts that were found near it, such as loom weights, bones, burnt seeds etc.

In order to extend the functionality of analyzing categorical-typological attributes, the 'Decoding Table' method was developed for breaking down these categorical definitions into shared elements. By doing this, categories with low frequencies could have a greater contribution to the analysis, and common traits of various categories could become meaningful. In addition, special programs were written to decipher the text variables of composite attributes. Many binary variables were created by these programs. Variables with low frequencies were dropped. The color attribute variables were converted to alternative color coding methods. Furthermore, various variables were created by calculating interactions amongst existing variables.

These processes resulted in a very wide data table of 1920 variables. This table was imported into Enterprise Miner and was further processed. Using this software it is easy to set and modify the various processes using a flow diagram. It also supplies automatic

tools for common Data Mining processes. First the metadata was defined, to enable further control for processing and analyzing the data. Using the SAS Code node various programs were written for the purpose of adding more computed variables. Special formulas for defining the presumed function of the vessel were developed using this node, and special programs were also written for dropping variables that tend to have a unimodal distribution. In other nodes missing values were imputed, rare categorical values were grouped, and variables with unsuitable values were rejected.

A group of Data Mining Neural Networks nodes were set to create prediction models for the various presumed functions and the temporary stratum number in which they were found. This actually created a set of fuzzy variables that defined the likeness of the vessel to the predicted trait. At the end of this process 1431 variables were left for use in the analysis.

The **fourth chapter** deals with the ways data mining techniques could contribute to the various typological aspects. It begins with an introductory summary of the history of the use and research of typology in archaeology. It is suggested that the goal of the typological process is **to compress a large number of items into representational categories in such a way that, on one hand, the loss of information will be minimal, and on the other hand, the level of detail will remain within the resources available for the publication.**

The partial success in creating an automatic typology of the Tell es-Safi bowls is summarized. The conclusion is that with existing software programs and methods, it will not be possible to produce a satisfying typology that will meet the suggested typology goals, and, as such, new algorithms need to be developed for this task. The tests that were done were based on using Cluster Analysis methods, which have a major disadvantage for this kind of task, in that they assign the same weight to all the variables in use.

Nevertheless some 'real' types can be discovered using these methods. The methods used included creating different sets of types for different sets of variables (morphological, finishing, fabric and contextual), and then using Cluster Analysis to produce composite types using the four sets of typologies. An example of a successful 'real' type that was discovered using this method is introduced. The importance of using

different types of attributes (not only morphological) in order to achieve an effective typology, is emphasized.

The results of the automatic morphological typology are compared with the typology constructed by Shai, which was done using an “intuitive” approach. This research found that some attributes, such as, strength of carination, direction of curvature above carination, and curvature of base, showed greater significance within the automatic typology using Cluster Analysis, than in the intuitive typology. In addition, using Decision Trees, further typical characteristics for the intuitively defined types were searched for, and in most attempts, were successfully found. These successes may indicate that these are ‘real’ types. Also found was that specific weight could be a good typological predictor.

The **fifth chapter** reviews some tests that were conducted according to pre-defined questions. These tests were implemented mainly by predicting certain attributes with Decision Trees. The results show the substantial potential inherent in using a table with many variables. With certain tests there was some difficulty reaching reasonable results, because of the sampling problems of the Tell es-Safi vessels. Using Decision Trees an attempt was made to characterize different classes of vessels, or traits, such as, vertical burnishing, reaction to hydrochloric acid, and reaction to a magnet. In addition, characterizations of the Late Bronze Age vessels from Area E were successfully done, resulting in a specific bowl type, with specific wall and rim angles, being revealed. The effectiveness of the fuzzy variables was also examined, by studying a spatial distribution map of vessels according to their likeness relation to serving vessels, and an interesting storage enclosure was found.

The **sixth chapter** reviews experiments that made use of Association Discovery Analysis, for the purpose of discovering different traits that tend to appear together. In order to do this there was a need to reformat the vessel table to suit this kind of analysis, which is not intended for finding associations that are meaningful. This program is designed mainly for market analysis, where the purpose is to find associations that may increase profit. Because of this, it was necessary to develop different ways of preparing the data so that ‘known in advance’ results would not show up. One of the solutions used

for solving this problem was to group the variables into classes of manufacturing phases, and to prevent the examination of association of traits of the same class.

The main weakness of the Association Discovery Analysis is in the handling of interval variables. This problem was partly solved by binning the values, and changing the variables levels to ordinal. Eventually there were many results, each one of which had to be carefully examined as to its true meaning.

The **seventh chapter** discusses the second case study with the data from Tel Batash. The purpose of this study was to examine methods for implementing Data Mining techniques on published data with a small number of variables. The main table contained 7809 observations of pottery rims that were found in Tel Batash. This table had a relatively small number of variables (about 31). For this reason the table was extended with contextual data and new variables which, with the help of Decoding Tables, were derived from the typological values. Each shard in this table was identified with a certain vessel type. It was therefore necessary to create a Decoding Table for vessel types, on the basis of the type description that appears in the published report. This process began by noting all the common typical attributes that are mentioned in this report, thus enabling formalization of a table with variables and values that matched the verbal descriptions. Furthermore, the basic morphological attributes from the representational drawings of each type, were measured. Another Decoding Table for fabric groups was created according to their verbal description. In addition, as was done with the Tell es-Safi data, Decoding Tables were created for certain typological-categorical values, such as burnish type or rim type.

In the next phase, the table was enriched with contextual data. Each locus had variables that indicated the existence of various finds within it. Variables for frequent words that appeared in the stratigraphic description were also created. In addition, using the feature definition, more calculated variables were created to indicate on which SFP (Site Formation Process) phase the locus is associated (building, occupation, destruction or abandonment). On the basis of the main shards table, new contextual variables that indicate which types of vessels were found in each locus or architectural unit, were created.

The results of the analysis of the Tel Batash data were much more meaningful than those obtained from the analysis of the whole vessels from Tell es-Safi. The reason for this is the greater chronological and spatial variability of the Tel Batash data. The main key to success in these kinds of analyses is in choosing the right sample. When contextual issues were examined, there was a need to use a sample of one vessel from each locus. When associations among different attributes were examined, there was a need to use a sample of vessel types. When fabric attributes were examined there was a need to use a sample of vessel types that had non-missing fabric data.

Decision Trees were used to characterize traits, such as, burnish type; or contextual values, such as, loci with bones. Chronological characterization of the vessels was done according to the stratum in which they were found. First the whole dataset was used, next, a sample of vessel types, and finally, a sample of types that also had fabric data.

Surprisingly, the analyses of the sample of vessel types produced very interesting results, in spite of its small size. Many of these results did not appear at all using the whole dataset analysis. This kind of data reflects the potter's function and production style trends, rather than distribution of use. An example of an interesting result from the analysis of this kind of sample, was the tendency of the absolute direction of the rim to turn outwards during the later stages of Iron Age II.

Association Discovery was also used with the Tel Batash data to discover traits that tend to appear together in the same vessel. The best results came from using the vessel types sample, because this kind of analysis, using the whole dataset, would result in artificial associations caused by the breaking down of types using the Decoding Tables. Another implementation of this analysis was done by examining vessel types that tend to appear in the same locus. It is interesting to note that the results of this kind of analysis demonstrated new aspects of the distribution of the examined types, aspects which were not mentioned in the published report.

To summarize the work, according to the new archaeological results that came out of the different analyses that were done in this research (summarized in index 9), it appears that there is a real potential in using the approach utilized in the present study. In addition, the use of data tables with many variables has also turned out to be worthwhile, but only on the condition that the right analysis methods are used, such as those that were

used in this research. It seems that in the near future, when inter-site databases will be created, the processing and analyzing of these databases, using Data Mining techniques, will be unavoidable. For this reason, a standard database structure should be decided upon and planned ahead to fit these needs.