Data Mining on an Ancient Attic Stele

Yuen Hsi Chang

Abstract

In 415 BC, prior to the Athenian expedition to Sicily, Andocides, among other Athenian elites, were testified for undermining the Eleusinian mysteries and mutilating the busts of sacred Hermae. Some of the convicted were condemned to death while others were sent into exile, and most of their possessions were seized and auctioned by the state to raise revenue. The final sales of these properties are inscribed on what is currently known as the "Attic Stele".

Archeologists have recovered many fragments of the Attic Stele, but a lot of information cannot be recovered as much of the inscriptions on the marble have faded over time. In this research project, Professor Dave Musicant, Professor Clara Hardy and I made various attempts in interpreting known information on the Attic Stele, so to gain a better understanding of the urban elites role in the ancient economy in the city of Athens.

1 Data Collection: After Clara translated and processed the data, we had a spreadsheet with 513 sales entries to work with. Each entry tells us up to eight pieces on information, but most of the rows are incomplete.

In the initial stele, the person to whom a property belongs to is recorded, but some names are missing due to the condition of the marble. Clara worked around this by specifying two "name" columns. In the first column, each person is given an ID, in which a new number is given each time its possible that there can be a new owner. If it were certain that the new ID indeed corresponds to a new person, a secondary column would specify that. These two columns are named "Person ID" and "New Person" respectively.

Information regarding the items sold was also shown. This includes the names of the sold items, the quantities and prices they were sold for, and the sales tax on them. In some entries, the exact price was not clear, but an approximate price range could be estimated by referencing the sales tax and the price of similar items. There were also times when only parts of the numerical value could be seen (e.g. if the word "thousand" is seen, the item could be anywhere between one thousand and ten thousand).

Finally, as the auctions didnt all occur at the same time, names are recurring on different stones, so there is also a column called "stelae #" and one called "column #" that differentiates between entries found on different stelae. As we are only working with a portion of the stelae that was recovered, our data only represents a portion of the sales that occurred during this event.

2 Data Recovery: Initially, the goal of this research project was to employ a recurring singular value decomposition technique to recover missing cells in the spreadsheet by looking at the distribution of data and drawing relationships between similar items. However, upon processing the spreadsheet, we found out that out of roughly fifty people, variations of items sold add up to 141. Out of the 141 items, 96 of them are uniquely owned by one person, and only 16 objects are owned by three or more people. On top of this,

many entries combine multiple items (e.g. an "amphora of wine" includes an amphora and some wine) as one entry. The algorithm to filling in missing data wouldnt work very well in this case, as its guesswork is highly dependent on locating similar entries.

In order to do any meaningful analysis over these people, they need to have many more items in common, so we combined different objects to form larger categories. We ended up grouping the 141 objects into six categories, namely food, containers, land-related, household tools/objects, furniture, and slaves. Unfortunately, these categories do not necessarily imply that the prices of objects among the same category are similar. For example, there is a person who had his entire estate auctioned, and another individual who sold some bricks. Both entries categorize as being land-related, but the prices of these objects differ by many degrees of magnitude. Data recovery using these given categories is not possible, as simply knowing that the item a person owns falls under a certain category does not feasibly narrow down the price he could have auctioned it for. Ultimately, we carried through with the project and performed data analysis with our given data without attempting to recover missing cells.

3 Assumptions: Our translated spreadsheet provides up to eight pieces of information about each entry, but we did not incorporate all the details for our data analysis. As we are aiming to better understand the economy of the Athenian elites, entries that are missing both the item name and the price of the item are discarded. Also, we simplified the problem at hand and didnt use the "'new person", "tax", "stele #", and "column #" columns. These details provide essential information if we are to perform guesswork regarding the person ID and item price further down the road, but it doesnt benefit us immediately for the purposes of understanding and interpreting the results.

4 Data Analysis: As previously stated, in order to make sense of the many variations of items sold in the auction, we categorized the items under six groups to work with. We did this in hopes of drawing relationships between item types and general welfare. Using a pivot table, we created a spreadsheet that displays the number of items each person sold under each category. Unfortunately, due to the diversity of the items being sold, knowing that someone has auctioned a large number of objects under a certain categories is not very telling of his objects aggregate value. For example, person 26 who has auctioned his full estate for 324,00 drachmas clearly sold for more value than person 18, who auctioned more than 10,000 vine poles for a total of 354 drachma. However, in the pivot table, person 18 is shown as someone auctioning over 10,000 land-related objects, whilst person 26 is recorded as someone who only auctioned one. This illustrates how it is difficult and not always feasible to draw connections between the aggregate value of items and the number of items being sold.

Furthermore, simply calculating the total revenue a person made from his/her auctions is troublesome as well. There are many entries that have illegible price columns, and also many entries that are missing item names and display just the price. Therefore, simply summing up the price column corresponding to each persons sales probably doesnt provide a thorough estimate regarding the aggregate values of their recorded items. The value could be off, and it would be very difficult to determine whether we overshot or undershot. To address this problem, we have two price columns for the pivot table. The first column only sums up the prices of entries for which the item name is known. This total doesnt account for entries with known item names and missing prices, but it doesnt deal with entries with known prices and unknown item names, so we would at least know that this column always undershoots. The secondary column sums up all the price fields visible, so each of its entries are always greater than the first price columns, and we could be overshooting or undershooting with this estimation. This column is not used to estimate the value of the known objects that were sold, but to give an idea on how much each persons auctions were worth.

Understanding our limitations, we carried through with the analysis and grouped similar people together, going by the types and number of objects each person owned. Unfortunately, probably due to a combination of the problems aforementioned, we didnt quite observe a relationship between a persons grouping and the aggregate value they auctioned their belongings for. Nevertheless, we still attempted to derive the economic situation in ancient Athens, treating the total value of each auction as a representation of the elites respective income.

5 Data Interpretation: The Gini coefficient is a number between zero and one that measures the inequality amongst people. A Gini coefficient of zero (perfect equality) means everyone has exactly the same income, and a Gini coefficient of one means that the richest 1% of the society makes 100% of all income (perfect inequality). Treating the value each person made in their respective auctions as a representation on income, we equated the Gini index of the people in Athens. With our data, it is calculated that the Gini coefficient in Athens was roughly 0.7, which is significantly higher (and hence more unequal) than even most developing countries. Chile, for example, had a gini coefficient of 0.526 in the late 2000s, before taxes were even accounted for (taxation usually pushes the coefficient down).

6 Limitations: Of course, it is a stretch assuming that the revenue from these auctions correspond to the income of the people. It could easily be problematic to associate the auctions to where the people stand economically, as simply because one owns more goods does not necessarily imply that they were required to auction correspondingly more items. On top of this, the data we are working with are only a fraction of the initial stele, so we are probably missing many significant auction entries. Finally, even the data set we are working with is missing a lot of cells and values it is highly possible that there were many important sales that we were unable to account for.

7 Further Work: Most of the excel work I laid down provides a strong basis for future work. If any Classics professor decides to add details to the spreadsheet by estimating prices and item names for respective entries, the pivot table would change accordingly, acting as a groundwork supporting future development. If I had more time to work on the data at hand, I would use a lot of the information we left out (see the "assumptions" section), and perhaps categorize the objects differently so they reflect the societal structure more accurately. It is possible that if more stelai are uncovered in the future, with better categories and more specific algorithms, we could even return to our initial plan involving the recovery of missing data through employing some data mining techniques.

During one of our meetings, we discussed the possibility of displaying the information we have on an interactive website, leaving out the data analysis. This could act as an interesting source for both Classics professors and student interested in the Attic Stele alike, if only to appreciate the history behind this ancient inscription, which was the aftermath of a large-scaled auction event that occurred in Athens prior the Hellenic Period.

8 Conclusion: Even though it has only been two weeks, the course of this research changed in many directions as we understood the data differently. I was initially naive to presume that this project is as simple as writing some java code to change the data in a form that R, a data analysis programming language can understand, and using its imputation library to fill in missing values. Upon understanding the data a little better, it turns out that too much data is missing, and too many categories are existent, for the algorithm to

tackle effectively. We ended up taking a completely different approach, performing data analysis rather than data recovery, and doing most of the work in Excel, without programming. This wasnt what I envisioned, but I have learnt a lot about various approaches, and I now have a greater appreciation regarding the broad spectrum of disciplines CS umbrellas. This project was interesting my biggest lesson I learnt from this project is that data is not always pretty to work with, but with appropriate tools and a little bit of background knowledge, one can always perform some level of data interpretation.