Article from

Predictive Analytics and Futurism

July 2016
Issue 13

Exploring the SOA Table Database

By Brian Holland

The Society of Actuaries database has historical values of several types of tables. This article goes through some basic data exploration techniques to show how different approaches look. Here I'm aiming for a quick view into the tables that are vectors, such as lapse rates by duration or ultimate mortality rates. We could also deal with matrices such as select and ultimate tables by laying rows out end to end to make a longer vector.

When would you do this type of thing in practice? You might if you had thousands of tables installed into a valuation system or pricing repository and you wanted to look for features. Those features could conceivably include typos, which should stand out and be caught.

There were 3,909 vectors among the 2,621 table files extracted from the SOA database. Some table files included two or more vectors. The winner was No. 1531,[1] which has 55 vectors of durational lapse rates by different segments of business. Of course, the rates could be organized differently than as a loose collection of vectors. However, the purpose here is to skip all organizational points and look quickly at the data as they are expressed in the database. Missing values are plugged with zero for that purpose, and different axes are lined up: durations in some cases, or ages in others. There are 141 dimensions: the longest vector has 127 values, but some only overlap, and the vectors go from 0 (like some attained ages) to 140 (a Brazilian mortality table[2]).

DIMENSION REDUCTION: WHAT IT IS

We are all intuitively familiar with some dimension reduction. Shadows reduce a 3-D object to 2-D; if the shadow is on a stick, the dimension drops from 3-D to 1-D. I find it helpful to imagine dimension reduction as rotation of a higher-dimensional object in a way to cast the widest shadow. The object does not have to just be three dimensions; here we reduce 141-dimensional objects to two dimensions. We will miss many facets of the data, but it is a start to get a view.

Singular value decomposition (SVD) is the dimension reduction technique used here. If we imagine each vector of the 3,909 in 141 dimensions, what we're doing with SVD is rotating the 141 axes so the first two rotated axes catch the biggest shadow, or dispersion of the points. There are many proper mathematical treatments on the web, which are good excuses to bone up on linear algebra. Note that principal component analysis (PCA) is closely related.

In Figure 1, each dot represents one of the 3,909 vectors. Around (0,0) we have most of the points; there are just a few outliers.

Figure 1: Plot of tables by first two right singular vectors (main dimensions)

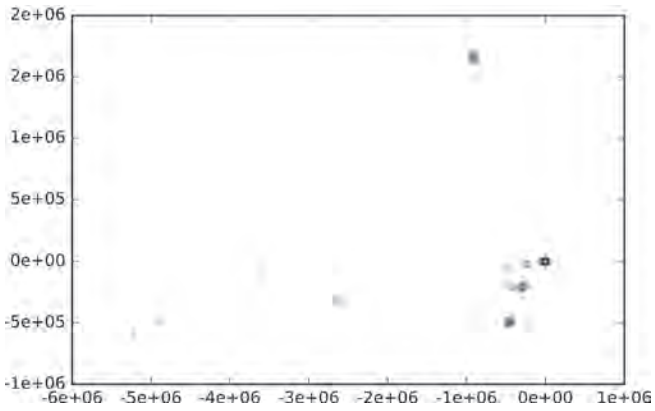
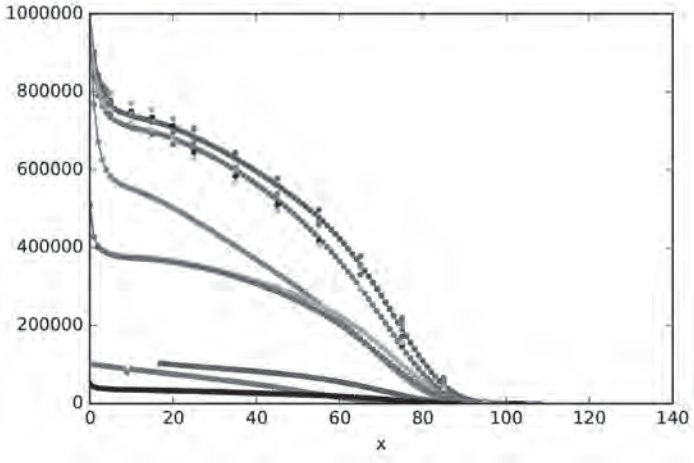


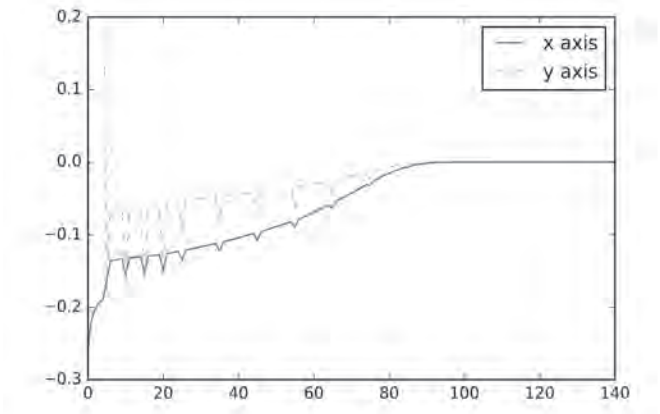
Figure 2 shows the vectors that the outlying points in Figure 1 represent. Those vectors are mostly English and Scottish life tables. It's no wonder they're outliers in Figure 1; they look nothing like mortality tables.

Figure 2: Actual main outlying vectors form Figure 1



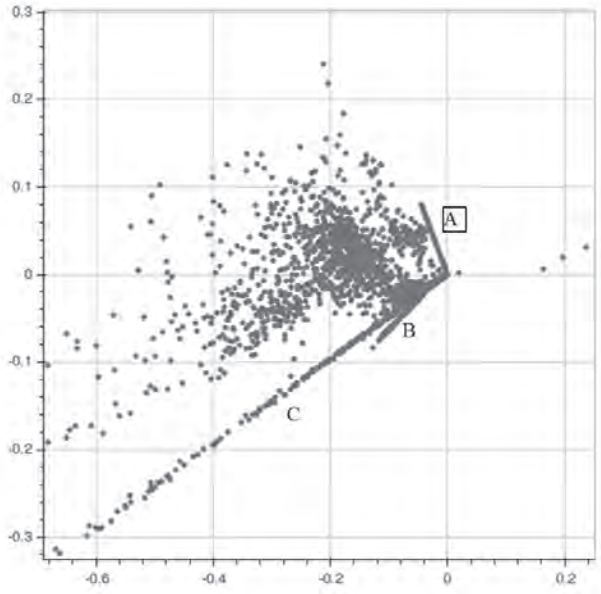
What do the axes in the graph above represent? Each axis is a certain level of each of the 141 values (dimensions), i.e., a vector as plotted below. To get back to the approximation of the original vector represented by one point in Figure 1, take the x and y coordinates, and use them to scale the x and y axis vectors in Figure 3.

Figure 3: Meaning of X and Y axes in Figure 1



Clearly, these vectors are a bit weird. There are regular dips. It turns out there are life tables with values only every several years, not every year, and those dominate the description of the data. What strikes me is that several patterns emerge anyway. Around (0,0) we have most of the vectors.

Figure 4: Zoomed in around 0,0: most (mortality, lapse, disability) vectors



Some structures jump right out. What turns out to be driving them is the areas that were missing.

A. Most points represent a vector from one of the truncated (ages 0 and 1) South American life tables.

B. Most points represent one of the South American life tables from 5-80.

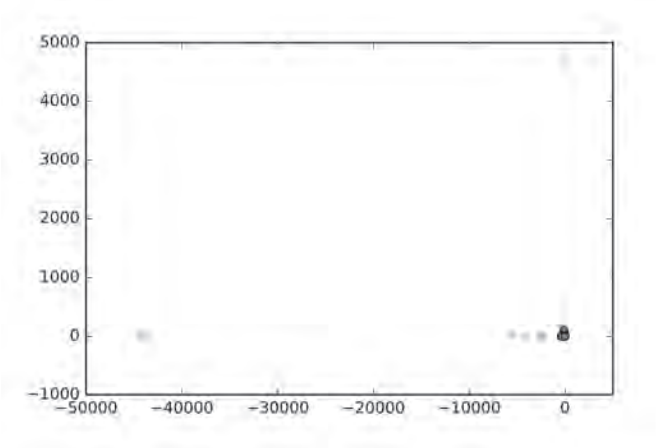
C. Most points represent disability tables or relative risk tables.

Omitting the English and Scottish life tables and others more than 200,000 from the origin (from eyeballing the graph), the

So what have we accomplished? In a quick analysis using a readily available algorithm, we've turned up an issue we can all relate to. ...

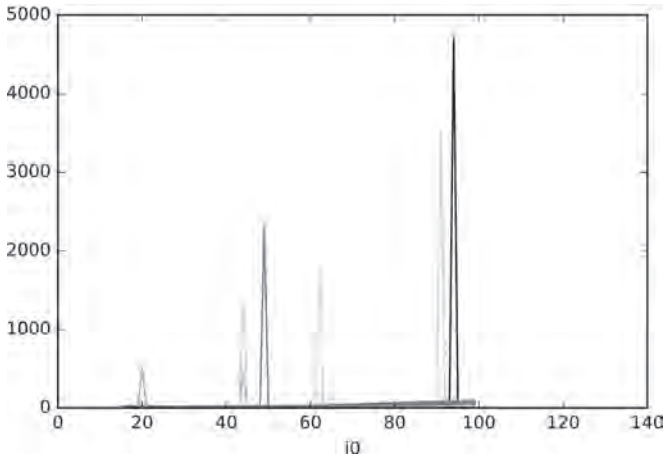
remaining tables would be plotted quite differently. There are some outliers along the y axis and some at about (-40000, 0). The former are mostly medical expense tables and the latter are more life tables. Both types are quite different from mortality rates. One of the medical expense tables is especially far off. By the way, browsing through these data I'm using a Python library called Bokeh, which allows easy browsing of large datasets. It can be told to show a text box when the mouse is over a point on the graph, which is how I tell the point's corresponding vector.

Figure 5: Decomposing again without main outliers



This outlier pointed me to some issues with scanned medical cost tables: Some values were missing decimals. The medical expense tables in question are from the 1970s and I doubt they are being used, but I'll still point it out to the table managers. That is exactly the kind of thing we are looking for. Grabbing those from the database and plotting them, we see some problems in the data entered for the 1974 medical expense tables. See Figure 6.

Figure 6: 1974 Medical Expense Tables (including typos)



Checking the values themselves, it's easy to see the decimal did not get typed or scanned for some values. To save paper, instead of printing them, I'll let you check them yourself, unless the database has been corrected by print time.

So what have we accomplished? In a quick analysis using a readily available algorithm, we've turned up an issue we can all relate to: an error in a valuation system. Are you ready to go through your own table repository?

Brian D. Holland, FSA, MAAA, is director and actuary, of Individual Life and A&H Experience Studies at AIG. He also serves as chair of the Predictive Analytics and Futurism Section Council. He can be reached at *brian.holland@aig.com.*

ENDNOTES

1 *http://mort.soa.org/ViewTable.aspx?&TableIdentity=1531.*

2 *http://mort.soa.org/ViewTable.aspx?&TableIdentity=2952.*