Snapshots of modern mathematics
from Oberwolfach

№ 12/2015

Visual Analysis of Spanish Male Mortality

J. S. Marron

Statistical visualization uses graphical methods to gain insights from data. Here we show how a technique called principal component analysis is used to analyze mortality in Spain over about the last hundred years. This data decomposition both reflects expected historical events and reveals some perhaps less expected trends in mortality over the years.

1 Data visualization

Statistical analysis is the study of data sets. Simple *data sets* can be thought of as (a collection of) tables which list variables and the values they get at data points (measurement points). For example, a data set can be a list of places and the temperature measured at these places on the 3rd of February, 1900, at 9 am. Or, it can be a list of all children in school and their height and age. A statistical analysis of these data sets employs techniques to gather understanding from the raw data – children of age of twelve years are more likely to be 150 cm tall than 190 cm tall[1]. An important, but all too often ignored, part of a statistical analysis is to look at – visualize – the data set. Common visualization methods can be in the form of graphs or charts. For modern complex data sets, it is not always clear how such visualization should be done, so this is an active research area.

[1] For more examples of statistical analysis see Snapshot 6/2014 "Statistics and dynamical phenomena" by H. Tong.

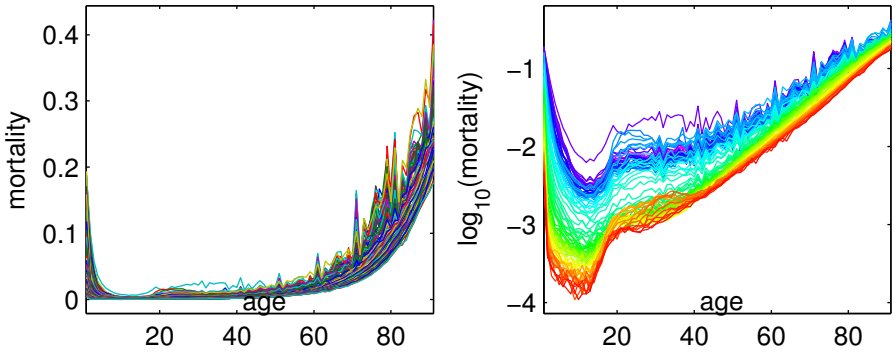

Figure 1: Raw mortality curves on the left, using mortality on the vertical axis, with standard rotating palate of colors. Log mortality is seen on the right to be the more natural scale, with rainbow color scheme indicating years. Shows overall age effects, plus long term trend towards lower mortality.

In this snapshot we will focus on *curves as data objects*, using the terminology of Object Oriented Data Analysis (OODA) coined by Wang and Marron [4]. See Marron and Alonso [1] for a detailed discussion of this idea, and also for further discussion of the data set studied here.

We will illustrate the main ideas of OODA data visualization using a *mortality* data set. Mortality reflects one's chance of dying, usually at some type of population level (for example, age). This is quantified, for a given group of people in a given time interval, as the fraction of the number who died out of the total number in the group. In our case the time intervals are the calendar years 1908–2002. In each year people are grouped according to ages from 0 to 90, and the mortality ratio is computed. Such a data set, for Spanish males, is studied in Figure 1. Each curve in the left panel corresponds to one year (1908–2002) and is a plot of mortality on the vertical axis as a function of age on the horizontal axis. Note that the different curves are distinguished by different colors. Most graphics software offer this as a default. Here the default palate of 7 colors from the Matlab software package was used to generate these graphics.

One limitation of the data view in the left panel is that much of the variation between curves is hard to see. This is because mortality ranges over several orders of magnitude, so it is hard to distinguish small values, especially in the childhood years when all curves appear to be essentially 0. A standard approach to this type of visualization challenge is to plot *logarithms* of the data. A logarithm of a number is the power to which another number (in our case, 10) needs to be taken to get that first number; for example, the logarithm of

100 is 2, as $10^2 = 100$. Since the differences between the curves are very small, smaller than 0.1, the logarithms of the death rates and their variations are relatively big (in the sense of absolute value) negative numbers. For example, at a point where the value of one curve was only 0.001 of that of another curve, in the logarithmic scale these two values are three units apart (as $0.001 = 10^{-3}$)[2]. Note how on the right panel, plotting the logs (short for logarithms) of the curves, variations at all ages are better revealed.

Another limitation of the mortality plot on the left is that it is hard to understand which year is which. This is overcome in the right panel of Figure 1 by using a different color scheme. Here a single color cycle is used for the entire time range 1908–2002, with colors following a rainbow color scheme with purple for 1908 through blue, cyan, green, yellow to red for the year 2002. This coloration already shows a clear trend: there has been a steady overall improvement in the mortality of Spanish males over the last century. This is due mostly to improvements in medicine and public health over that time range.

2 Curves as data objects

Until now, the data we analyzed – our data set – was comprised of death-rates, ages and years. We visualized this data as curves in various ways as to better understand it. Let us take a step further, and see if we can glean more information, by regarding the curves themselves as data to be analyzed. Thinking of the curves as a data set, it is natural to think about the center point. The left panel of Figure 2 shows the curve which is obtained by plotting the mean at every point (along the horizontal axis) of the curves in the right panel of Figure 1. Note that this mean curve shows expected human life-cycle patterns. The far left is high because it is dangerous to be an infant. After that mortality drops rapidly through childhood, then gradually grows, as older people have a proportionally higher chance of dying.

A perhaps unexpected feature is a series of small peaks. Note that these peaks are not occurring at random times, and instead are equally spaced. Furthermore, they appear only at decade ages (multiples of 10). This is because these peaks are an artifact of poor record keeping in the earlier part of the time range. In the earlier years, when an older person died, there was sometimes uncertainty as to the precise age, so there was some rounding in the reported age. This is clear from the peak at decade ages, with valleys in both the immediately preceding and following ages.

[2] Incidentally, this technique is also very useful when dealing with very large variations of very large positive numbers. By regarding the logarithms of very large quantities we get smaller numbers (we exchange 100 for 3, say), and so get the data into a manageable size; for such a use, see Figure 3 in Snapshot 5/2015 *Chaos and chaotic fluid mixing* by T. Solomon.

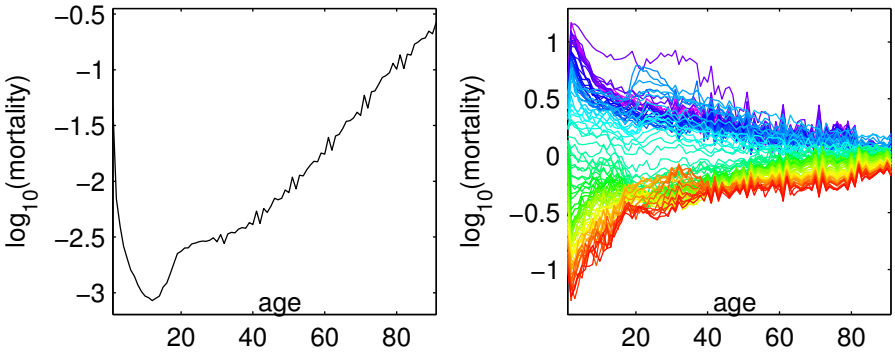

Figure 2: Mean (center) curve on the left, containing main age effects. Mean residuals on the right showing overall improving mortality and no age effects.

While the mean curve is interesting, additional insights are available from careful consideration of the variation about the mean. A simple view is shown in the right panel of Figure 2, which is the *mean residuals*. These curves are just the data curves from the right of Figure 1, with the mean in the left of Figure 2 subtracted. As expected, the overall improvement (drop) in mortality is strongly apparent in these residual curves. Also note that the main age impacts on mortality are essentially missing, showing that these are overall average effects which have not noticeably changed over time.

A more refined view of the data comes in the left-hand panel of Figure 3, the *first principal component* (PC1) loading plot. This is a technique to highlight major differences by filtering out lesser ones.

To understand what we do, it is helpful to think of a space of curves. This would be a space with dimensions for all possible variables: one dimension for each age-group. For example, in Figure 1 we had one dimension for ages and one for mortality, so each point in that (two-dimensional) space represented the mortality at an age-group. Such spaces, where we are interested in the values of points according to different dimensions (called *coordinates*) are called *vector spaces*, and the points in vector spaces are called *vectors*. In our new space each point (vector) will represent an entire mortality curve; that is, each point encodes the mortality at all ages at one year. Now that we have a space of curves, we mark the mean-centered residuals from the right panel of Figure 2 in this space and find the direction of maximal variance[3]. Each data point

[3] Simply put, variance measures how much a data-set is spread out.

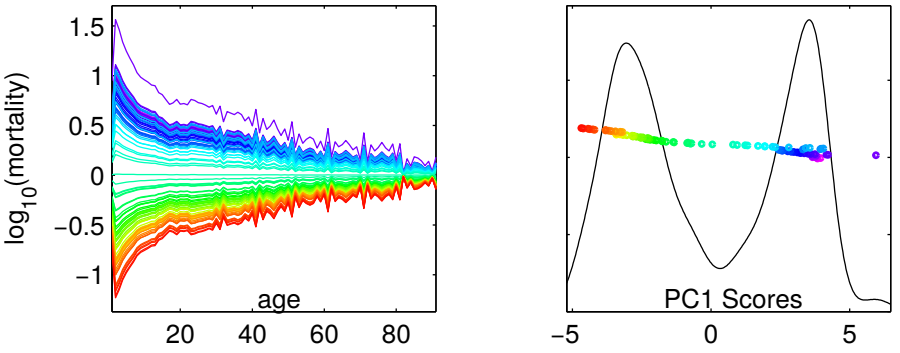

Figure 3: PC1 Loadings plot on the left, showing overall improvement is dominant mode of variation. PC1 scores on the right show the dominant mode of variation.

(mean-residual curve) can be projected on this direction vector. This projection can be thought of as measuring where the shadow of each point falls on the line that marks the direction of maximal variation. The resulting curves, all of them represented as points lying on the same line, will be multiples of the common direction vector curve. This can be seen on the left panel of Figure 3, where the projected points (again as mortality curves) are multiples of the same shape (negative multiples are mirror images). In a sense, this procedure reveals the variation among the data points in the most varied aspect. Both the magnitudes of the projected curves (called *scores*) and the shapes of the curves give useful insights.

The common shape of the curves reflects the expected fact that overall health improvements benefit essentially the entire population. However, the amount of improvement is a decreasing function of age, because the positive effects of medical technology wane with age. Note also that the decade age blips that appear in the mean curve on the left in Figure 2 are also important features here. This time they go upwards in the earlier years, showing the age rounding was stronger in earlier years. The blips go downwards later, but this is because the age effect disappeared later, and here we look at the difference with the mean.

A deeper look at the scores, can be found in the right panel of Figure 3[4]. These are the coefficients of the 1st PC projection. Each dot corresponds to a mortality curve, where the horizontal coordinate marks where it lies on the

[4] The dark plot at the background of the right panel (and in Figure 4) represents *density estimation*. It shows the probability of a curve to have a certain score-value.

maximal variance vector. Since we are dealing with the residual curves, the mean curve is the zero point. This means the farther the score-value of a dot is from zero, the farther the curve is from the mean – the more it contributed to the mean, positively or negatively. The colors of the dots correspond to each year, and they are arranged vertically in year order – the earlier years are lower and the later ones are higher. The bluish purple on the far right is 1918. It has the highest positive score, meaning it contributed most to the mean. That was the year of the perhaps most important epidemiological event in world history. Soldiers returning from World War I carried a horrible strain of flu that killed millions world wide. The large death toll in Spain that year is reflected by the position of this dot. The fact that this year is an outlier is apparent even in the raw data plot (see the same shade of magenta curve) in the right of Figure 1. After that, there were some overall improvements, until the next swing to the right. Some might guess that was World War II, but in fact Spain was not a combatant in that war. Actually, the light blue dots correspond to the late 1930s when there was a terrible civil war fought in Spain. After that there has been a steady shift to the left, especially as overall health conditions have improved over time.

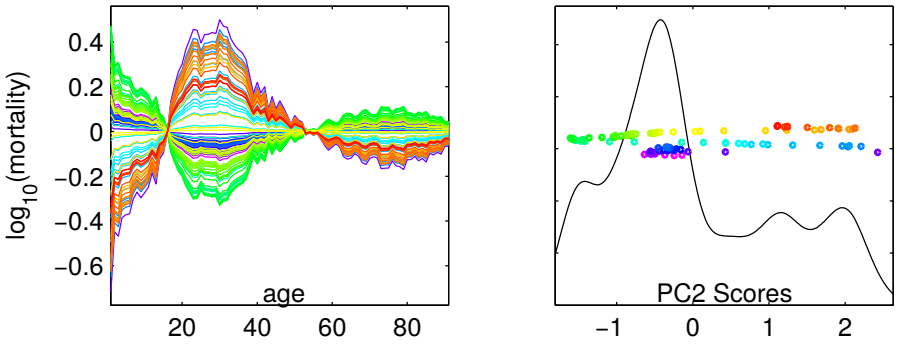

Figure 4: PC2 Loadings on the left.

Figure 4 gives a deeper look at how mortality has changed over time. This time the focus is on the second principal component (PC2). The procedure is the same as in the PC1 case, only now we look at a direction of maximal variation which is perpendicular to the direction we found for the first principal component. In a sense, this procedure ignores the variation of the first direction in order to emphasize the next most prominent variation. The loadings plots are again all multiples of a single curve and are color-coded by year. Now the color pattern is less easy to interpret. However an important point for the pattern of curves is that this direction is about the difference between the group of 20–45

year old males and the union of the young and old.

While it is not so clear from the color pattern on the left, some clear trends do emerge from the scores plot on the right (same format as Figure 3). Again the year 1918 (bluish purple) and the Spanish Civil War (light blue) are very prominent, because the 20–45 year old males were dying at a relatively greater rate during these years. Then there was rapid improvement up to the mid 1950s (green), when things began to get worse (not overall, but for the 20–45 year old males). During this era, automobiles became commonly available, and the tendency for unsafe driving in this demographic led to steadily increasing death rate. That trend was fortunately reversed during the early 1990s (orange–red) with the advent of seat-belts and other car safety features, as well as much improved road design.

These examples show the power of principal component analysis for the decomposition of complex data sets of curves into more easily interpretable pieces. For much more on this type of analysis, often called *Functional Data Analysis*, see Ramsay and Silverman [2, 3].

Image credits

All images created by the author for this publication.

References

[1] J. S. Marron and A. M. Alonso, *Overview of object oriented data analysis*, Biometrical Journal **56** (2014), 732–753.

[2] J. O. Ramsay and B. W. Silverman, *Applied functional data analysis: methods and case studies*, Springer Series in Statistics, Springer, New York, 2002.

[3] _____, *Functional data analysis*, Springer Series in Statistics, Springer, New York, 2005.

[4] H. Wang, J. S. Marron, et al., *Object oriented data analysis: Sets of trees*, The Annals of Statistics **35** (2007), no. 5, 1849–1873.

J. S. Marron *is a professor of Statistics and Operations Research at the University of North Carolina, Chapel Hill, USA*

Mathematical subjects
Numerics and Scientific Computing, Probability Theory and Statistics

Connections to other fields
Finance, Humanities and Social Sciences

DOI
10.14760/SNAP-2015-012-EN

Snapshots of modern mathematics from Oberwolfach are written by participants in the scientific program of the Mathematisches Forschungsinstitut Oberwolfach (MFO). The snapshot project is designed to promote the understanding and appreciation of modern mathematics and mathematical research in the general public worldwide. It started as part of the project "Oberwolfach meets IMAGINARY" in 2013 with a grant by the Klaus Tschira Foundation. The project has also been supported by the Oberwolfach Foundation and the MFO. All snapshots can be found on www.imaginary.org/snapshots and on www.mfo.de/snapshots.

Junior Editor
Daniel Kronberg
junior-editors@mfo.de

Senior Editor
Carla Cederbaum
senior-editor@mfo.de

Mathematisches Forschungsinstitut
Oberwolfach gGmbH
Schwarzwaldstr. 9–11
77709 Oberwolfach
Germany

Director
Gerhard Huisken