Exploratory functional data analysis via PCA

Rob J Hyndman

23 December 2024

Qu et al. (2025) have produced a fascinating paper on the tools that are available for exploratory analysis of functional data. Much of the literature has focused on statistical models for functional data, and related theory, so it is great to see the important pre-modelling work receiving some attention.

Amongst the methods they describe, several use functional principal component decomposition (Ramsay & Dalzell 1991) to transform the functional data into a lower-dimensional space. Then some standard EDA tools are applied to the first few principal component scores, and the results translated back into the original functional space. For example, this was the approach used in the functional bagplot and functional HDR boxplot proposed in Hyndman & Shang (2010). While there is no guarantee that the features of interest that are present in the original functional data will be preserved in the PCA space, in practice this almost always leads to useful results. As well as providing some helpful visualization tools, this approach can also be used for anomaly detection, giving an alternative approach to those methods based on statistical depth that are discussed by Qu et al. (2025).

Figure 1 (left) shows US mortality rates between 1933 and 2022, obtained from the Human Mortality Database (2024). Each line denotes the mortality rates as a function of age for one year, with the colors in rainbow order corresponding to the years of observation. Overall, we see a large decrease in mortality rates during early childhood years, then an increase during teenage years. After about age 30, the rates increase almost linearly on a log scale. Comparing the curves over time, we see that the rates have steadily fallen for all ages up to about 95 years, with more than a 10-fold reduction in mortality rates at around age 10.

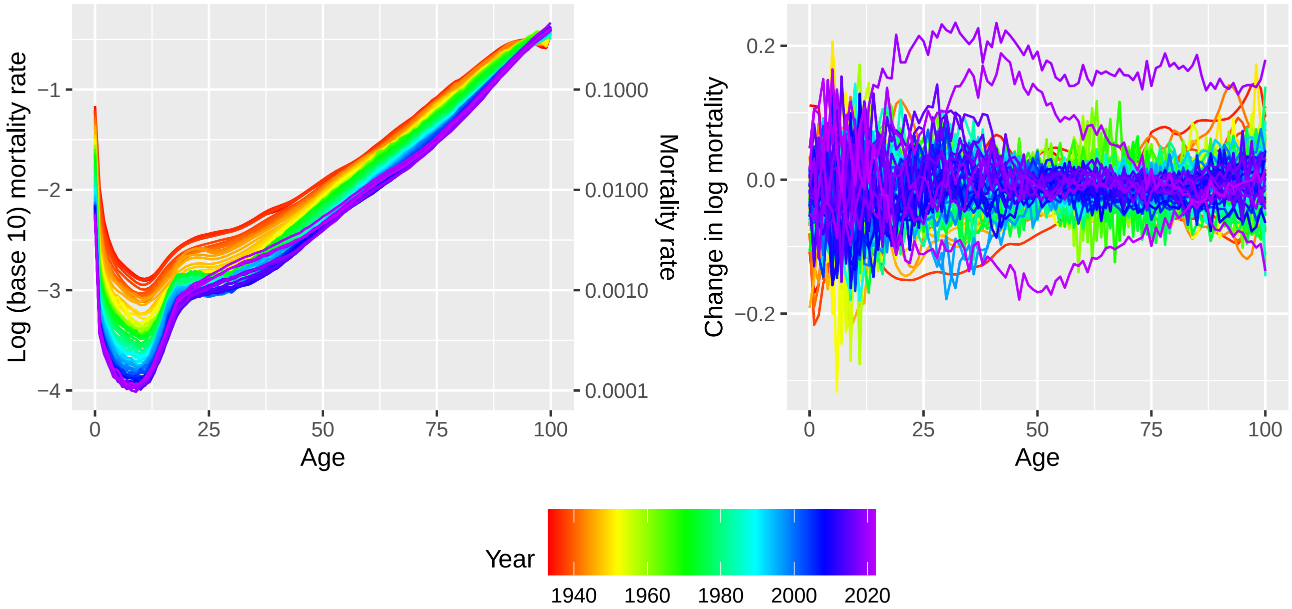

Figure 1: *Left: US age-specific mortality rates for 1933–2022. Right: Annual differences in log mortality rates.*

Clearly the data are non-stationary due to the steady decline over time, so we consider the differences in the log mortality rates over time, as shown in right panel. Now several functional observations stand out as having different behaviour from the others, including three (in purple) from the last few years of data.

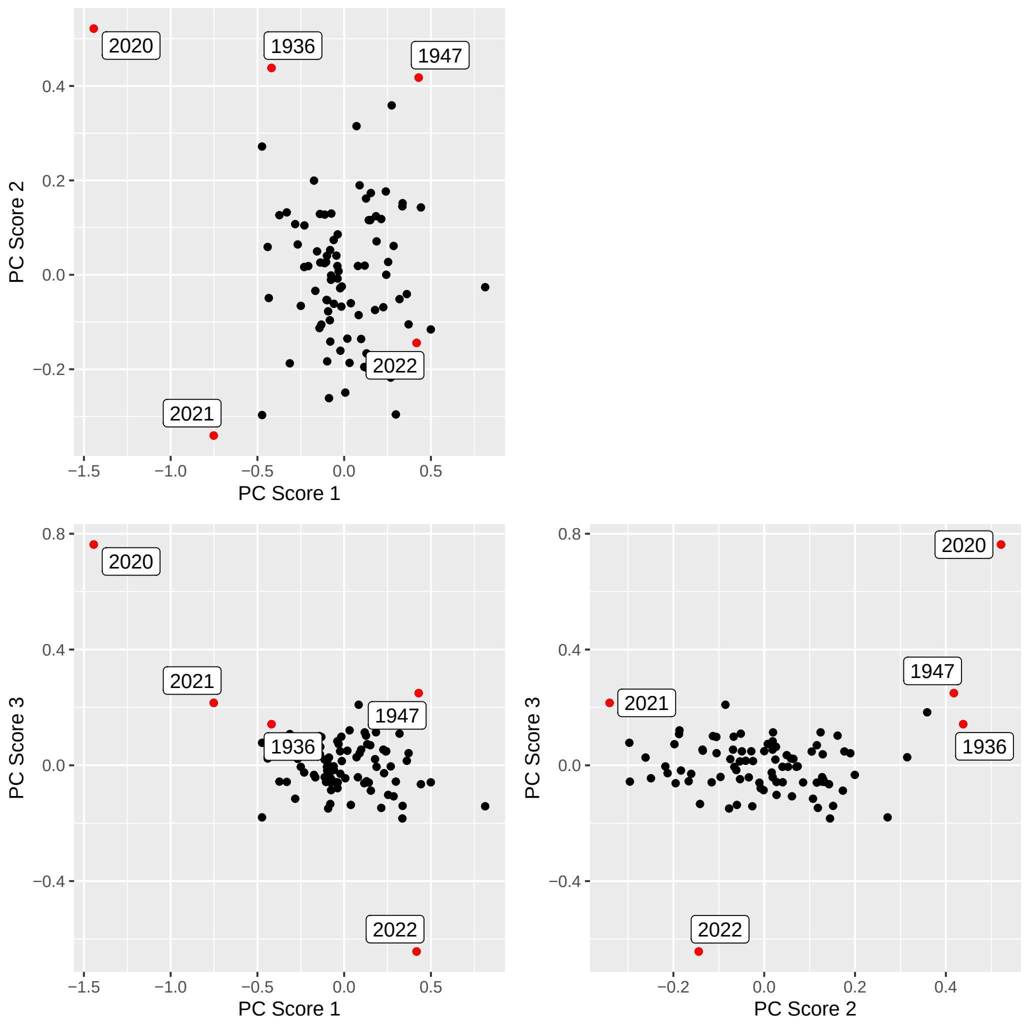

Figure 2: *Pairwise scatterplots of the first three principal component scores for the US age-specific log mortality annual differences.*

We will use principal component scores to detect anomalies in this data set. Because we are interested in anomaly detection, we do not want the principal component decomposition to be affected by the anomalies we are trying to detect. Consequently, we will use the robust principal component method proposed by Croux, Filzmoser & Oliveira (2007), applied to the annual differences in the log mortality rates (shown on the right of Figure 1), to obtain the first three principal component scores. These are shown in Figure 2. The loadings (not shown) suggest that the first PC corresponds to ages 0–40, the second PC increases with age after age 30, while the third PC contrasts children under 10 with people above age 25.

The lookout anomaly detection algorithm (Kandanaarachchi & Hyndman 2022) has been applied to the first three PC scores. This estimates a multivariate kernel density estimate of the 3-dimensional data set, and fits a generalized Pareto distribution (GPD) to the top 10% of the most extreme “surprisal” values (equal to minus the log of the estimated density at each observation). Those points with probability less than 0.5 under the GPD are labelled in Figure 2 (giving an effect false positive rate of 5%). The last three years of data (2020–2022) are identified as anomalies (probably due to COVID-19), along with 1936 (at the end of the Great Depression, Tapia Granados & Diez Roux (2009)) and 1947 (due to rapid improvement in mortality after WW2). Note that all three principal component scores are needed to identify these anomalies.

War deaths are excluded from the data set, as they took place outside the country, so the war years are not seen as anomalous.

Figure 3 shows the years identified as anomalous against the backdrop of all other years in the data set. While the last three years stand out from the rest of the data, the data for 1936 and 1947 are not so obviously anomalous from the data plot alone.

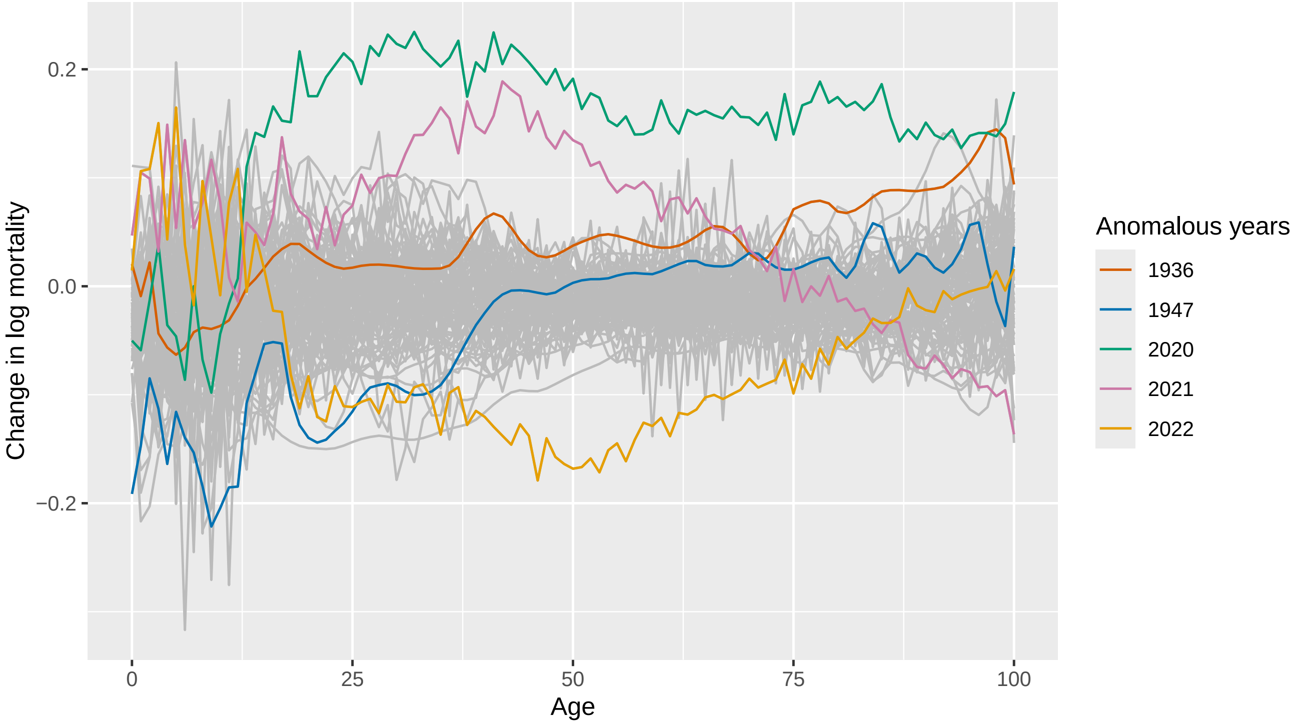

Figure 3

When the "directional outlyingness" method of Dai & Genton (2019) is applied to these data, only 1947, 2020, 2021 and 2022 are identified as anomalies. The increase in mortality at the end of the Great Depression is missed, although it is arguably more anomalous than 1947 (especially after age 70) which is identified.

This general approach to exploratory functional data analysis, using the PCA space rather than the original functional data space, can be employed for many other standard data analysis tasks such as assessment of data quality, identifying trends and seasonality, change point detection, density estimation, feature engineering, and more. The advantage is that it allows immediate use of the vast array of multivariate data analysis tools that already exist, rather than having to develop new tools for functional data. It provides a familiar and computationally efficient set of tools that is complementary to those that work more directly in the functional data space.

The code to reproduce the results in these comments is available at https://github.com/robjhyndman/EFDA.

References

Croux, C, P Filzmoser & M Oliveira (2007). Algorithms for Projection-Pursuit Robust Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **87**, 218–225.

Dai, W & MG Genton (2019). Directional outlyingness for multivariate functional data. *Computational Statistics & Data Analysis* **131**, 50–65.

Human Mortality Database (2024). Data downloaded on 22 December 2024. Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France). www.mortality.org.

Hyndman, RJ & HL Shang (2010). Rainbow plots, bagplots and boxplots for functional data. *J Computational & Graphical Statistics* **19**(1), 29–45.

Kandanaarachchi, S & RJ Hyndman (2022). Leave-one-out kernel density estimates for outlier detection. *J Computational& Graphical Statistics* **31**(2), 586–599.

Qu, Z, W Dai, C Euan, Y Sun & MG Genton (2025). Exploratory functional data analysis. *TEST*. in press.

Ramsay, JO & C Dalzell (1991). Some tools for functional data analysis. *J Royal Statistical Society: Series B* **53**(3), 539–561.

Tapia Granados, JA & AV Diez Roux (2009). Life and death during the Great Depression. *Proceedings of the National Academy of Sciences* **106**(41), 17290–17295.