Eine visuelle Analyse der Sterblichkeit männlicher Spanier

J. S. Marron, ist Professor der Statistik und Unternehmensforschung an der University of North Carolina in Chapel Hill, USA

Zusammenfassung

Die statistische Visualisierung benutzt graphische Methoden um Erkenntnisse aus Daten zu gewinnen. Wir zeigen wie mit dem Verfahren der Hauptkomponentenanalyse die Sterblichkeit in Spanien im Laufe der letzten hundert Jahre analysiert werden kann. Diese Datenzerlegung zeigt sowohl erwartete geschichtliche Ereignisse auf, als auch einige, teilweise überraschende Entwicklungen der Sterblichkeit im Laufe der Zeit.

1 Datenvisualisierung

Die statistische Analyse beschäftigt sich mit Datensätzen. Einfache *Datensätze* bestehen im Grunde aus (einer Ansammlung von) Tabellen, die eine Liste von Variablen mit ihren dazugehörigen Werten an bestimmten Datenpunkten (Messpunkten) enthalten. Ein Datensatz kann beispielsweise eine Liste von Orten und die Temperatur, die am dritten Februar 1900 um 9 Uhr an ihnen gemessen wurde, enthalten. Oder es handelt sich um eine Liste aller Schüler, die eine bestimmte Schule besuchen, zusammen mit ihrer Körpergröße und ihrem Alter. Eine statistische Analyse solcher Datensätze nutzt Verfahren um aus den gegebenen Rohdaten Erkenntnisse zu gewinnen, z.B. dass zwölfjährige Kinder eher 150 cm als 190 cm groß sind.[1]. Ein sehr wichtiger, aber leider allzu oft vernachlässigter Teil der statistischen Analyse besteht darin, sich den Datensatz anzuschauen – zu visualisieren. Gebräuchliche Visualisierungsverfahren zeigen sich in Form von Graphen und Diagrammen. Weil bei modernen, komplexen Datensätzen nicht immer ganz klar ist wie sie visualisiert werden sollen, handelt es sich hierbei um ein aktives Forschungsgebiet.

In diesem Schnappschuss beschäftigen wir uns hauptsächlich mit *Kurven als Datenobjekten* und verwenden Begriffe aus der objektorientierten Datenanalyse (OODA, Object Oriented Data Analysis), welche von Wang und Marron [4] geprägt wurde. Eine ausführliche Abhandlung dieses Konzepts und eine weiterführende Diskussion des hier untersuchten Datensatzes findet sich bei Marron und Alonso [1].

Anhand eines Datensatzes, der Daten zur Sterblichkeit enthält, stellen wir die Hauptkonzepte der OODA-basierten Datenvisualisierung vor. Unter Sterblichkeit verstehen wir die Wahrscheinlichkeit dafür, dass eine Person bei einer bestimmten Populationsstufe (z.B. in einem bestimmten Alter) stirbt. Bei einer gegebenen Gruppe von Leuten in einem gegebenen Zeitraum wird die Sterblichkeit quantifiziert indem die Anzahl der Verstorbenen durch die Gesamtgröße der Gruppe geteilt wird. In unserem Beispiel besteht der beobachtete Zeitraum aus den Kalenderjahren 1908 bis 2002. Für jedes dieser Jahre werden die Menschen nach Alter von 0 bis 90 gruppiert und die entsprechende Sterblichkeitsrate berechnet. Ein solcher Datensatz, der für den männlichen Anteil der spanischen Bevölkerung erhoben wurde, ist in Abbildung 1 wiedergegeben. Jede der Kurven im linken Diagramm entspricht dabei einem Jahr (1908–2002) und ist als Graph der Sterblichkeit dargestellt, welche auf der vertikalen Achse aufgetragen ist und als Funktion des auf der horizontalen Achse aufgetragenen Alters interpretiert wird.

[1] Weitere Beispiele der statistischen Analyse gibt es im Schnappschuss 6/2014 "Statistics and dynamical phenomena" von H. Tong.

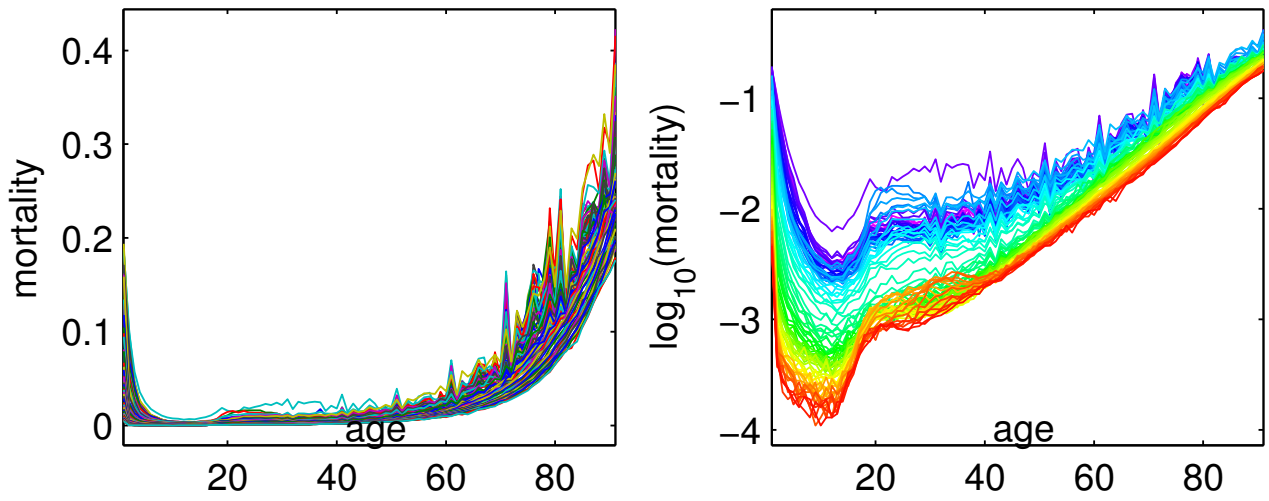

Abbildung 1: Links: Unbearbeitete, mit einer Standardpalette eingefärbte Sterblichkeitskurven mit der Sterblichkeit (mortality) auf der vertikalen Achse. Rechts: Logarithmierte Sterblichkeitskurven. Die logarithmische Skalierung wirkt natürlicher, die einzelnen Jahre werden durch die Farben des Regenbogens unterschieden. Diese Kurven zeigen nicht nur allgemeine, altersbedingte Effekte, sondern auch einen auf sinkende Sterblichkeit hinweisenden Langzeittrend.

Die Kurven der verschiedenen Jahre unterscheiden sich in ihrer Farbe – ein Feature, das die meisten Visualisierungsprogramme von Haus aus anbieten. In unserem Fall wurde die aus sieben Farben bestehende Standardpalette des Matlab-Softwarepakets verwendet, mit dem die obigen Graphiken erzeugt wurden.

Die Darstellung unserer Daten im linken Diagramm wird dadurch eingeschränkt, dass ein Großteil der Variation zwischen den Kurven nur schwer erkennbar ist. Das liegt daran, dass die Werte der Sterblichkeit sich über mehrere Größenordnungen erstrecken und es somit schwierig wird kleine Werte zu unterscheiden – vor allem in den Kindheitsjahren, wo alle Kurven sich im Wesentlichen bei 0 befinden. Ein gebräuchlicher Ansatz dieses Visualisierungsproblem zu lösen ist den *Logarithmus* der Daten darzustellen. Der Logarithmus einer Zahl ist diejenige Potenz zu der eine andere Zahl (in unserem Fall die 10) erhoben werden muss, so dass sich dadurch die erste Zahl ergibt; beispielsweise ist der (Zehner-)Logarithmus von 100 die 2, denn $10^2 = 100$. Weil die Unterschiede zwischen den Kurven sehr klein sind – kleiner als 0,1 – werden die Logarithmen der Sterberaten und ihre Unterschiede betragsmäßig (also im Sinne des Absolutbetrags) zu relativ großen, negative Zahlen. An einer Stelle, an welcher der Wert eine Kurve beispielsweise nur das $0,001$-fache einer anderen Kurve beträgt, unterscheiden sich die beiden Kurven bei einer logarithmischen Skalierung um drei Längeneinheiten (denn $0,001 = 10^{-3}$)[2]. Dieser Effekt zeigt sich im rechten Diagram, wo die Logs (Abkürzung für Logarithmen) der Kurven dargestellt sind und sich die Unterschiede bei jedem Alter deutlicher abzeichnen.

Eine weitere Einschränkung des linken Sterblichkeitsdiagramms besteht darin, dass wir kaum erkennen können welche Kurve zu welchen Jahr gehört. Dieses Problem wird im rechten Diagramm von Abbildung 1 durch den Einsatz einer anderen Farbpalette beseitigt. Hier wird ein einziger Farbzyklus für die Jahre 1908 bis 2002 verwendet, dessen Farben ein Regenbogenschema durchlaufen. Dieser beginnt für das Jahr 1908 bei blau, durchläuft die Farben cyan, grün und gelb und färbt schließlich die Kurve für das Jahr 2002 rot ein. Dieses Farbschema zeigt einen klaren Trend: im Laufe des letzten Jahrhunderts gab es eine stetige, allgemeine Besserung (d.h. eine Senkung) der Sterblichkeit in der männlichen, spa-

[2]Nebenbei bemerkt ist diese Herangehensweise auch dann sehr nützlich, wenn starke Variationen zwischen großen, positiven Zahlen vorliegen. Dadurch, dass der Logarithmus aus sehr großen Zahlen wesentlich kleinere macht (wir ersetzen mit ihm z.B. 100 durch 3), können wir die Werte eines Datensatzes wesentlich überschaubarer machen; dieser Trick wird beispielsweise in Abbildung 3 des Schnappschusses 5/2015 *Chaos and chaotic fluid mixing* von T. Solomon verwendet.

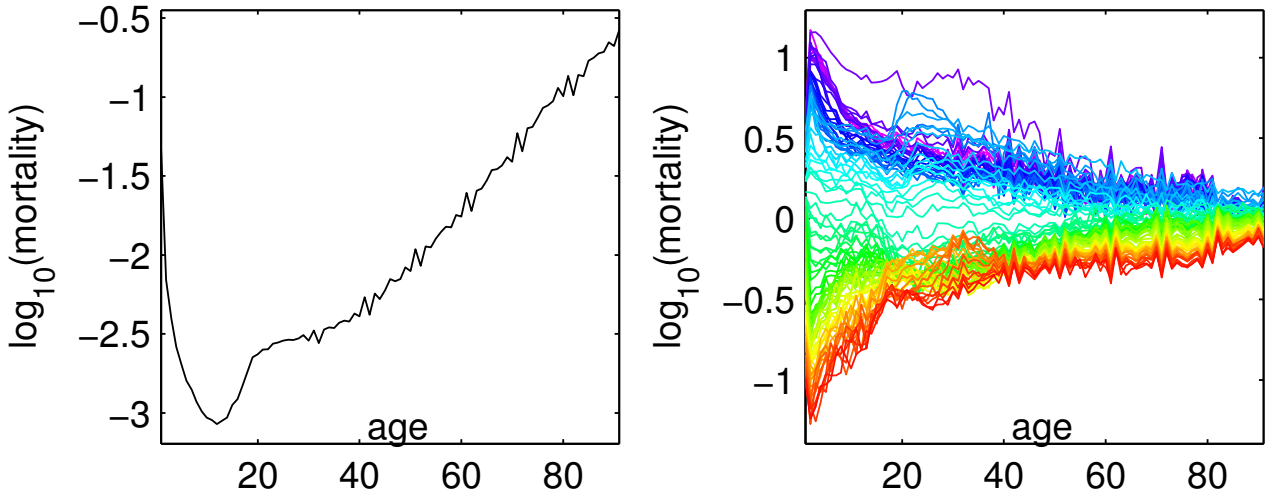

Abbildung 2: Links: Die Mittelwertskurve mit den allgemeinen Alterseffekten. Rechts: Residualkurven, welche die allgemeine Verbesserung der Sterblichkeit zeigen und keine Alterseffekte aufweisen.

nischen Bevölkerung. Das liegt vor allem an dem Fortschritt der Medizin und der Verbesserung des Gesundheitswesens in diesem Zeitraum.

2 Kurven als Datenobjekte

Bis jetzt bestanden die Daten, die wir analysierten – unser Datensatz – aus Sterblichkeitsraten, Altersklassen und Jahren. Um diese Daten besser zu verstehen, haben wir sie auf verschiedene Arten visualisiert. Wir gehen nun einen Schritt weiter und möchten wissen, ob wir noch mehr Informationen gewinnen können, indem wir die Kurven selbst als zu analysierende Daten auffassen. Wenn wir die Kurven als Datensatz auffassen, liegt es nahe ihre Mittelpunkte zu betrachten. Das linke Diagramm von Abbildung 2 zeigt die Kurve, welche wir erhalten, wenn wir an jeder Stelle (der horizontalen Achse) den Mittelwert aller Kurven im rechten Diagramm von Abbildung 1 berechnen. An dieser Mittelwertkurve erkennen wir den erwarteten, menschlichen Lebenszyklus. Ganz links ist die Kurve hoch, denn es ist gefährlich ein Säugling zu sein. Danach, in den Kindheitsjahren, fällt die Sterblichkeit rasant ab um über die Lebensjahre hinweg schrittweise anzusteigen: schließlich haben ältere Menschen im Verhältnis zu jüngeren ein höheres Todesrisiko.

Etwas überraschend wirken möglicherweise die kleinen Spitzen. Diese tauchen jedoch nicht zu zufälligen Zeitpunkten auf, sondern kommen in gleichen Abständen vor. Außerdem zeigen sie sich nur bei dekadischen Altern (Vielfachen von Zehn). Das liegt daran, dass diese Spitzen ein Überbleibsel schlechter Buchhaltung in den früheren Zeitabschnitten unseres Gesamtzeitraums sind. Wenn in früheren Zeiten ein alter Mensch starb, war sein genaues Alter manchmal ungewiss. Deswegen wurde in solchen Fällen das amtliche Sterbealter einfach gerundet. Das zeigt sich klarerweise in den Spitzen bei den dekadischen Alterszahlen und den unmittelbar davor und danach liegenden Tälern.

Auch wenn die Mittelwertskurve schon für sich interessant ist, ergeben sich zusätzlich Erkenntnisse, wenn wir genauer auf die Variation um den Mittelwert achten. Eine einfache Darstellung dieser Variation ist im rechten Diagramm von Abbildung 2 gegeben, nämlich durch die *Residuen (um den Mittelwert)*. Diese Kurven gehen von den Datenkurven aus dem rechten Diagramm von Abbildung 1 hervor nachdem die Mittelwertskurve aus dem linken Diagramm von Abbildung 2 von ihnen abgezogen wurde. Wie erwartet ist die allgemeine Verbesserung (Senkung) der Sterblichkeit deutlich in diesen Residualkurven zu erkennen. Zudem sind die oben erwähnten altersbedingten Auswirkungen auf die Sterblichkeit verschwunden, was zeigt, dass sie allgemeine Effekte sind, die sich nicht merklich mit der

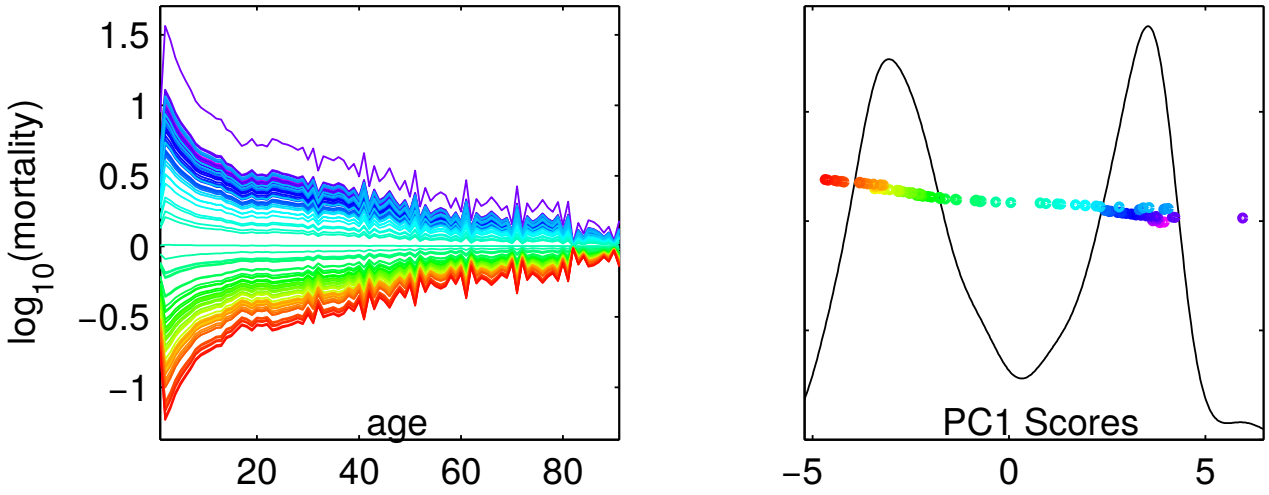

Abbildung 3: Links: Der Loadings-Plot für die erste Hauptkomponente (PC1). Dieser zeigt, dass die allgemeine Verbesserung der Sterblichkeit die Hauptursache der Datenvariation darstellt. Links: Der dazugehörige Score-Plot der ersten Hauptkomponente (PC1), der die Hauptursache der Variation genauer beschreibt.

Zeit verändern.

Eine genauere Betrachtung der Daten ist im linken Diagramm von Abbildung 3 zu sehen. Dieses Diagramm wird Loadings-Plot der *ersten Hauptkomponente* (PC1, *principal component 1*) genannt. Bei diesem Verfahren werden starke Unterschiede hervorgehoben indem die weniger starken herausgefiltert werden.

Um dieses Vorgehen zu verstehen ist es hilfreich über den Raum der Kurven nachzudenken. Dieser ist ein Raum, der für jede Variable – jede Altersklasse – eine Dimension besitzt. In Abbildung 1 haben wir beispielsweise eine Dimension für die verschiedenen Altersklassen und eine für die Sterblichkeit. Somit ist jeder Punkt in diesem (zweidimensionalen) Raum durch die Sterblichkeit einer bestimmten Altersklasse festgelegt. Räume, in denen wir uns für die Werte der Punkte in ihren verschiedenen Dimensionen (welche *Koordinaten* genannt werden) interessieren, heißen *Vektorräume*. Ihre Punkte nennen wir *Vektoren*. In unserem neuen Vektorraum repräsentiert jeder Punkt (Vektor) eine ganze Sterblichkeitskurve; das bedeutet, dass jeder Punkt die Sterblichkeit aller Altersklassen in einem bestimmten Jahr kodiert. Jetzt, wo wir einen Raum für unsere Kurven haben, markieren wir in ihm die Residuen um den Mittelwert aus dem rechten Diagramm von Abbildung 2, um die Richtung der maximalen Varianz[3] zu bestimmen. Jeder Datenpunkt (d.h. jede Residualkurve) kann nun auf diesen Richtungsvektor projiziert werden. Wir können uns diese Projektion wie folgt vorstellen: wir messen wo der Schatten jedes Punktes auf die Gerade in Richtung der maximalen Variation fällt. Die resultierenden Kurven, die als Punkte repräsentiert auf einer gemeinsamen Gerade liegen, werden Vielfachen dieses gemeinsamen Richtungsvektors sein. Das zeigt sich im linken Diagramm von Abbildung 3, wo die projizierten Punkte (wieder als Sterblichkeitskurven) Vielfachen der gleichen Form sind (wobei negative Vielfache sich wie Spiegelbilder verhalten). Dieses Verfahren offenbart gewissermaßen die Variation zwischen den Datenpunkten im Hinblick auf die größte Variationsursache. Sowohl die Längen der projizierten Punkte (ihre *Scores*) als auch die Formen der Kurven liefern nützliche Erkenntnisse.

Die gemeinsame Gestalt der Kurven spiegelt auch die erwartete Tatsache wieder, dass Verbesserungen des allgemeinen Gesundheitswesens im Grunde der ganzen Bevölkerung nutzen. Allerdings schrumpft die Höhe der Besserung bei ansteigendem Alter, denn auch die positiven Auswirkungen des medizinischen Fortschritts lassen gerade bei steigenden Lebensalter nach. Auch in dieser Betrachtung sind die Zacken bei den dekadischen Altersgruppen, die wir schon in der Mittelwertskurve auf der linken Seite

[3]Einfach ausgedrückt, misst die Varianz wie stark die Werte eines Datensatz streuen.

von Abbildung 2 beobachteten, ein wichtiges Merkmal. Diesmal gehen sie jedoch in den frühen Jahren (des Gesamtzeitraums) nach oben, was zeigt, dass das Runden des Sterbealters gerade in diesen frühen Jahren häufiger vorkam. Später zeigen die Zacken nach unten, was daran liegt, dass dieser Alterseffekt später verschwunden ist und wir hier ja immer die Differenz zum Mittelwert betrachten.

Eine genauere Betrachtung der Scores ist im rechten Diagramm von Abbildung 3 wiedergegeben[4]. Diese sind die Koeffizienten der Projektion auf die erste Hauptkomponente. Jeder Punkt entspricht einer Sterblichkeitskurve, wobei die horizontale Koordinate anzeigt, wo genau sie auf dem Vektor in Richtung der maximalen Varianz liegt. Weil wir mit den Residualkurven arbeiten, entspricht hier die Mittelwertkurve dem Nullpunkt. Je weiter also der Score eines Punktes von der Null entfernt ist, desto höheren Abstand hat die entsprechende Kurve von der Mittelwertkurve – und umso mehr hat sie zu den Mittelwerten beigetragen, sei es positiv oder negativ. Die Farben der Punkte entsprechen wieder den Jahren. Außerdem sind sie genau in ihrer Reihenfolge vertikal angeordnet – die früheren Jahre befinden sich unten, die späteren oben im Diagramm. Der lila-blaue Punkt ganz rechts steht für das Jahr 1918. Er besitzt den höchsten positiven Score, was bedeutet, dass er am meisten zum Mittelwert beigetragen hat. In diesem Jahr trat das vielleicht wichtigste, epidemiologische Ereignis der Weltgeschichte ein. Die Soldaten, die aus dem ersten Weltkrieg zurückkehrten, brachten einen furchtbaren Abkömmling der Grippe mit sich, der Millionen von Menschen auf der ganzen Welt tötete. Die hohe Todeszahl in Spanien in diesem Jahr spiegelt sich durch die Position dieses Punktes wieder. Dass es sich bei diesem Punkt um einen Ausreißer handelt, ist auch schon aus der Visualisierung der Rohdaten im rechten Diagramm von Abbildung 1 erkennbar (dargestellt durch die magentafarbene Kurve). In den Jahren danach gab es allgemeine Verbesserungen, die bis zum nächsten Schwung nach rechts anhielten. Es wäre denkbar, dass es sich hierbei um eine Auswirkung des zweiten Weltkrieg handelt, aber tatsächlich war Spanien in diesem Jahr gar nicht an den Kampfhandlungen des Krieges beteiligt. Stattdessen repräsentieren die hellblauen Punkte die späten dreißiger Jahre, als in Spanien ein grausamer Bürgerkrieg ausgetragen wurde. Danach rücken die Punkte wieder nach links, vor allem nachdem sich das allgemeine Gesundheitswesen im Laufe der Zeit verbessert hatte.

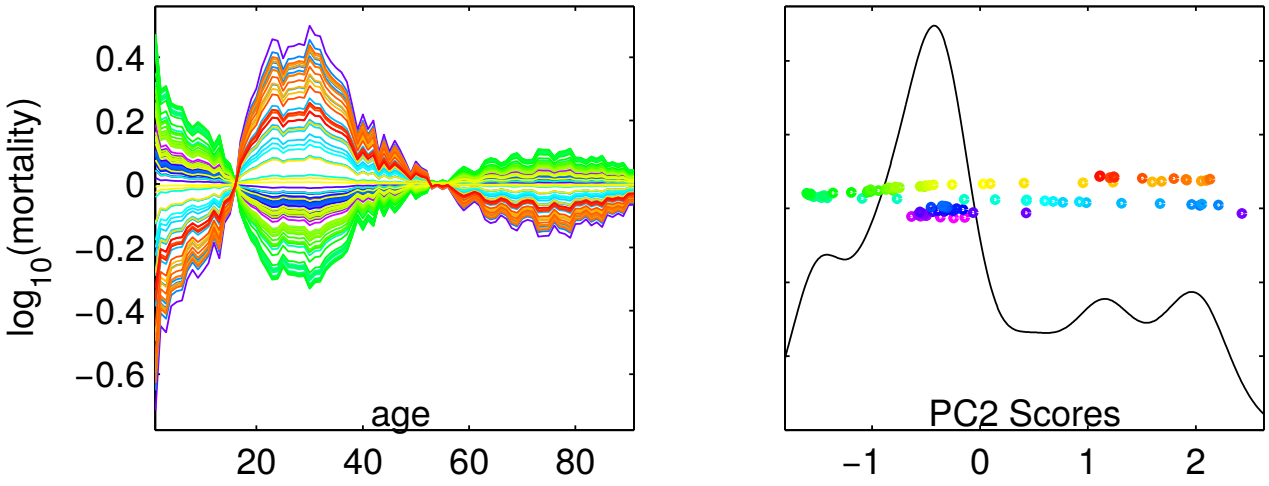

Abbildung 4: Links: Der Loadings-Plot für die zweite Hauptkomponente (PC2). Rechts: Der dazugehörige Score-Plot.

Abbildung 4 zeigt noch genauer, wie sich die Sterblichkeit über die Jahre veränderte. In dieser Abbildung wird die zweite Hauptkomponente (PC2) betrachtet. Das entsprechende Verfahren gleicht demjenigen für die erste Hauptkomponente (PC1), nur dass wir jetzt denjenigen Richtungsvektor der maximalen Variation betrachten, der senkrecht auf der Richtung steht, die wir bei der ersten Hauptkomponente verwendeten. Dabei wird gewissermaßen die Variation der ersten Richtung ignoriert um

[4]Die dunkle Kurve im Hintergrund des rechten Diagramms (und in Abbildung 4) ist die *geschätzte Dichte*. Sie zeigt wie wahrscheinlich die verschiedenen Score-Werte für die Kurven sind

die nächstgrößere Variation hervorzuheben. Die graphische Darstellung der Loadings besteht wieder aus den Vielfachen einer einzelnen Kurve, wobei ihre Farben die entsprechenden Jahre kodieren. In diesem Fall ist das Farbschema weniger leicht zu interpretieren. Das Hauptmerkmal an dieses Kurvenmusters ist jedoch, dass diese Richtung die Differenz zwischen der Gruppe der 20- bis 45-jährigen Männer und der übrigen, aus allen jüngeren und älteren bestehenden Gruppe, darstellt.

Während die Farbmuster im linken Diagramm kein ganz klares Bild liefern, zeigen sich im dazugehörigen Scores-Plot auf der rechten Seite (was dasselbe Format wie in Abbildung 3 besitzt) einige deutliche Entwicklungen. Wieder treten das Jahr 1918 (lila) und der Spanische Bürgerkrieg (hellblau) deutlich hervor, weil die 20 bis 45 Jahre alten Männer während dieser Jahre in vergleichsweise höherer Zahl verstorben sind. Danach ergibt sich eine Besserung bis zu der Mitte der fünfziger Jahre (grün). In den Fünfzigern stieg die Sterblichkeit erneut (nicht für alle, aber gerade für 20- bis 45-jährige Männer). Zu dieser Zeit wurden Automobile erschwinglich und die Neigung zu gefährlichem Fahrverhalten in gerade dieser demographischen Gruppe führte zu einer stetig wachsenden Sterblichkeitsrate. Glücklicherweise kehrte sich dieser Trend in den frühen Neunzigern (orange bis rot) dank der Einführung von Sicherheitsgurten und anderen Sicherheitsmaßnahmen im Verkehr, sowie verbesserter Straßenplanung, wieder um.

Diese Beispiele zeigen, wie effektiv die Hauptkomponentenanalyse komplexe Datensätze von Kurven in einfacher zu interpretierende Teile zerlegen kann. Mehr über dieses Analyseverfahren, oft *funktionelle Datenanalyse* genannt, gibt es bei Ramsay und Silverman [2, 3].

Acknowledgement

Schnappschüsse moderner Mathematik aus Oberwolfach werden von Teilnehmerinnen und Teilnehmern des wissenschaftlichen Programms des Mathematischen Forschungsinstituts Oberwolfach (MFO) geschrieben. Das Schnappschuss-Projekt hat zum Ziel, Verständnis und Wertschätzung für moderne Mathematik und mathematische Forschung in der allgemeinen Bevölkerung weltweit zu fördern. Es begann als Teil des Projekts „Oberwolfach trifft IMAGINARY“, welches von der Klaus Tschira Stiftung gefördert wird. Das Projekt wurde auch von der Oberwolfach Stiftung sowie vom MFO unterstützt. Alle Schnappschüsse können unter `www.imaginary.org/snapshots` sowie unter `www.mfo.de/snapshots` abgerufen werden. Der Orginaltext kann unter `http://www.mfo.de/math-in-public/snapshots/files/visual-analysis-of-spanish-male-mortality` gefunden werden, inklusive aller Abbildungen in Farbe.

Mathematische Gebiete:
Numerik und Wissenschaftliches Rechnen, Wahrscheinlichkeitstheorie und Statistik

Verbindung zu anderen Gebieten:
Finanzwissenschaften, Geistes- und Sozialwissenschaften

Übersetzt aus dem Englischen:
Daniel Katona

Editor:
Daniel Kronberg - junior-editor@mfo.de

Chefeditorin:
Carla Cederbaum - senior-editor@mfo.de

Literatur

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

[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] J. O. Ramsay and B. W. Silverman. *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(5):1849–1873, 2007.