Vanella, Patrizio

Working Paper

Stochastische Prognose demografischer Komponenten auf Basis der Hauptkomponentenanalyse

Hannover Economic Papers (HEP), No. 597

Provided in Cooperation with:
School of Economics and Management, University of Hannover

Suggested Citation: Vanella, Patrizio (2017) : Stochastische Prognose demografischer Komponenten auf Basis der Hauptkomponentenanalyse, Hannover Economic Papers (HEP), No. 597, Leibniz Universität Hannover, Wirtschaftswissenschaftliche Fakultät, Hannover

This Version is available at:
https://hdl.handle.net/10419/172851

Stochastische Prognose demografischer Komponenten auf Basis der Hauptkomponentenanalyse

Patrizio Vanella

Gottfried Wilhelm Leibniz Universität Hannover

Center for Risk and Insurance

Demographic and Insurance Research Center

pv@ivbl.uni-hannover.de

14. Juni 2017

Zusammenfassung

Für eine adäquate Prognose der zukünftigen Bevölkerung auf Basis von Kohorten-Komponenten-Methoden ist eine geschlechter- und altersspezifische Betrachtung erforderlich, da ansonsten die zukünftige Struktur der Bevölkerung nicht korrekt bestimmt werden könnte. Da altersspezifische demografische Größen untereinander allerdings hochkorreliert sind und zusammen einen hochdimensionalen Komplex bilden, bedarf es einer Methodik, die sowohl die Korrelationen zwischen den Zufallsvariablen einbezieht, als auch die effektive Dimension des Prognoseproblems verringert. Die Hauptkomponentenanalyse dient beiden Zwecken simultan.

Ziel dieses Beitrages ist die an Anwender aus dem Bereich der Bevölkerungswissenschaften gerichtete Vorstellung der Hauptkomponentenanalyse aus mathematisch-statistischer Sicht. Zudem wird auf Grundlagen der Zeitreihenanalyse eingegangen, die für eine korrekte stochastische Prognose unerlässlich sind. Die Anwendung wird anhand der Prognose ausgewählter alters- und geschlechtsspezifischer Mortalitäts- und Fertilitätsraten inklusive Prognoseintervallen für Deutschland, Italien und Österreich illustriert.

Schlagworte: Quantitative Bevölkerungswissenschaften, Multivariate Verfahren, Prognostik

1 Einführung und Motivation

Offizielle Bevölkerungsprognosen werden häufig auf Basis von deterministischen Kohorten-Komponenten-Modellen durchgeführt (vgl. z.B. Statistisches Bundesamt 2015: 13). Im Vergleich zu deterministischen Ansätzen sind stochastische Prognosen zu bevorzugen (vgl. Keilman und Pham 2000: 42-43), da diese neben der Prognose des wahrscheinlichsten Szenarios unendlich viele mögliche Szenarien mit zugehörigen Wahrscheinlichkeiten identifizieren und quantifizieren können. Stochastische Modelle lassen sich ebenfalls auf Basis der Komponenten Fertilität, Migration und Mortalität aufbauen. Im Rahmen dieses Beitrages wird ein Ansatz vorgestellt, der altersspezifische Prognosen für die Komponenten Fertilität und Mortalität ermöglicht. Dabei werden nicht nur einzelne Szenarien quantifiziert, sondern neben den c.p. wahrscheinlichsten Szenarien ebenfalls Prognoseintervalle durch Computersimulationen erzeugt. Da bei den Komponenten Korrelationen zwischen den unterschiedlichen Altersjahren und Geschlechtern sowie zwischen den einzelnen Zeitpunkten einer Zeitreihe identifiziert werden können, müssen bei Zukunftsprognosen Auto- sowie Kreuzkorrelationen berücksichtigt werden.

Im Rahmen dieser Ausarbeitung erfolgt daher eine kurze Vorstellung der Hauptkomponentenanalyse, wobei der Fokus auf der Erklärung ihres Nutzens und einer kurzen Darstellung ihrer Funktionsweise liegt. Zudem wird auf in der praktischen Anwendung wichtige Konzepte der Zeitreihenanalyse eingegangen, wobei sich auf die für die Anwendung notwendigen Aspekte beschränkt wird. Dieser Beitrag sollte daher als Leitfaden in statistischen Ämtern oder auch Forschungsinstituten verstanden werden, die konkreten Prognosen dienen lediglich illustrativen Zwecken und sollen nicht als Prognosen der tatsächlichen zukünftigen Entwicklung verstanden werden. Die Erläuterung der Methodik steht an dieser Stelle im Mittelpunkt. Der Fokus des Beitrages liegt in der Implementierung der dargestellten statistischen Konzepte zur Modellierung und Prognose der Komponenten der demografischen Entwicklung. Dabei wird die Anwendung anhand der Prognose ausgewählter altersspezifischer Fertilitätsraten für Deutschland, Italien und Österreich illustriert, wobei die Modelle ebenfalls Prognosen für andere europäische Länder ermöglichen und die Methodik problemlos auf alters- und geschlechtsspezifische Mortalitätsraten übertragen werden kann. Als Anwendungsbeispiel werden Prognosen für 80-jährige Männer für die drei genannten Länder illustriert.

2 Einführung in die Hauptkomponentenanalyse

Bevölkerungsprognosen werden in der Regel über die Kohorten-Komponenten-Methode nach Geschlecht und Alter differenziert durchgeführt. Dabei wird der Bevölkerungsstand am Ende des Vorjahres jährlich um die Todesfälle und die Auswanderungen reduziert sowie um die Geburten und Einwanderungen erhöht, die im aktuellen Jahr anfielen (vgl. Weidner et al. 2015: 372). Eine adäquate Prognose der zukünftigen Bevölkerung sollte daher idealerweise auf Basis einer statistischen Modellierung und Vorhersage der altersspezifischen Fertilitätsraten (ASFR), altersspezifischen Mortalitätsraten (ASMR), altersspezifischen Immigrationsraten (ASIR) und der altersspezifischen Emigrationsraten (ASER) durchgeführt werden. Diese Größen sind jedoch zum Teil untereinander hochkorreliert. Weiterhin führt die Vielzahl an zu prognostizierenden Variablen zu einer sehr hohen Dimensionalität. Auf diese beiden Probleme müssen Anwender mit angemessenen Methoden reagieren. In diesem Rahmen empfiehlt sich die Hauptkomponentenanalyse, welche die zwei angeführten Probleme simultan behandelt.

Die Idee der Hauptkomponentenanalyse ist eine orthogonale Transformation der ursprünglichen Variablen in ebenso viele neue, unkorrelierte Variablen, welche als Hauptkomponenten (HK) bezeichnet werden. Die Methode eignet sich dabei besonders in Situationen, in denen die Originalvariablen nicht, wie bei Regressionsanalysen üblich, kausal miteinander in Verbindung gebracht werden sollen (vgl. Chatfield und Collins 1980: 57). Daher ist die Hauptkomponentenanalyse besonders für Prognosen altersspezifischer Kennziffern im demografischen Kontext geeignet. Jede HK stellt dabei eine Linearkombination aller N Originalvariablen dar. Sei $F_{i,t}$ die i-te ASFR in Periode t. Dann berechnet sich die j-te HK $P_{j,t}$ in der gleichen Periode aus (vgl. Chatfield und Collins 1980: 58):

$$P_{j,t} = \sum_{i=1}^{N} e_{i,j,t} F_{i,t} =: \overrightarrow{e_{j,t}^{T}} \overrightarrow{F_t} \qquad (1)$$

Dabei kann $e_{i,j,t}$ als eine Art Korrelationskoeffizient zwischen der i-ten ASFR und der j-ten HK interpretiert werden.

Die HK werden im Rahmen der Hauptkomponentenanalyse absteigend nach der Variation gebildet, die sie aus den Originalvariablen erklären. D.h. die erste HK erklärt den größten Anteil an der Varianz aus den ursprünglichen Variablen. Wie bereits angesprochen, liegt darin der zweite große

Vorteil der Hauptkomponentenanalyse. Durch die Transformation kann in der Regel ein komplexes System mit vielen Variablen effektiv auf wenige Dimensionen reduziert werden, weil die ersten HK bereits den Großteil der Varianz erklären (vgl. Chatfield und Collins 1980: 57-58). Das wird an späterer Stelle am Beispiel europäischer ASFR verdeutlicht.

An dieser Stelle ist wichtig, darauf einzugehen, wie die HK berechnet werden. Wie bereits erwähnt, wird die erste HK so gewählt, dass sie so viel Variation der ASFR wie möglich erklärt. Statistisch bedeutet dies, dass die Kovarianz der ASFR unter Anpassung der Koeffizienten maximiert werden muss. Die Korrelation zwischen Hauptkomponente und ASFR wird ab diesem Punkt vereinfachend als zeitinvariant angenommen, sodass der Index t weggelassen werden kann. Wird die Kovarianzmatrix von $\boldsymbol{F}$ mit $\boldsymbol{\Sigma}$ bezeichnet, dann gilt für die Varianz der ersten HK:

$$Var[P_1] = Var\left[\overrightarrow{e_1^T}\boldsymbol{F}\right] = \overrightarrow{e_1^T}\boldsymbol{\Sigma}\overrightarrow{e_1} \tag{2}$$

Der Vektor $\overrightarrow{e_1}$ kann dabei willkürlich gewählt werden. Um eine eineindeutige Lösung des Maximierungsproblems zu erhalten, muss jedoch eine Nebenbedingung für die Elemente von $\overrightarrow{e_1}$ hinzugefügt werden. Es lässt sich zeigen: Die Bedingung, der Vektor habe die Länge 1, sorgt dafür, dass die Transformation genau orthogonal wird. Ein Vektor hat die Länge 1, wenn das Skalarprodukt des Vektors mit sich selbst den Wert 1 ergibt, d.h. (vgl. Handl 2010: 120-121):

$$\overrightarrow{e_1^T}\overrightarrow{e_1} = 1 \tag{3}$$

Nach der Methode von Lagrangia lassen sich die stationären Punkte einer Funktion $f(\vec{x})$ unter der Nebenbedingung $g(\vec{x}) = c$ auffinden, indem die zugehörige Lagrange-Funktion (auch Lagrangiana genannt) $\mathcal{L}(\vec{x}, \lambda)$ bzgl. ihrer stationären Punkte[1] untersucht wird. Die Lagrangiana ist wie folgt definiert[2]:

$$\mathcal{L}(\vec{x}, \lambda) = f(\vec{x}) - \lambda[g(\vec{x}) - c] \tag{4}$$

Die stationären Punkte der Lagrangiana werden dadurch aufgefunden, dass die partiellen Ableitungen erster Ordnung nach $\vec{x}$ und λ mit null gleichgesetzt und anschließend nach $\vec{x}$ und λ aufgelöst werden, sodass ein eineindeutiges Gleichungssystem zu lösen ist (vgl. Glaister 1984: 184-189):

[1] Dies können entweder lokale Minima, lokale Maxima oder Sattelpunkte sein.

[2] Grds. lässt sich eine Lagrangiana auf unendlich viele Nebenbedingungen erweitern. An dieser Stelle wird sich allerdings auf den Fall einer Nebenbedingung beschränkt, weil dies für Hauptkomponentenanalyse hinreichend ist.

$$\begin{aligned}
\frac{\partial \mathcal{L}(\vec{x},\lambda)}{\partial x_1} &= \frac{\partial f(\vec{x})}{\partial x_1} - \lambda\left[\frac{\partial g(\vec{x})}{\partial x_1}\right] = 0,\\
&\vdots\\
\frac{\partial \mathcal{L}(\vec{x},\lambda)}{\partial x_n} &= \frac{\partial f(\vec{x})}{\partial x_n} - \lambda\left[\frac{\partial g(\vec{x})}{\partial x_n}\right] = 0,\\
\frac{\partial \mathcal{L}(\vec{x},\lambda)}{\partial \lambda} &= c - g(\vec{x}) = 0
\end{aligned} \tag{5}$$

Dies wird an einem einfachen Beispiel illustriert. Es sei $f(x_1, x_2, x_3) = x_1 + 2x_2 + 3x_3$ eine dreidimensionale Funktion, die unter der Nebenbedingung $x_1 + x_2^2 + x_3^2 = 1$ maximiert werden soll. Folglich wird die Lagrangiana aufgestellt:

$$\mathcal{L}(\vec{x},\lambda) = x_1 + 2x_2 + 3x_3 - \lambda[x_1 + x_2^2 + x_3^2 - 1]$$

Diese wird folglich auf ihre stationären Punkte untersucht.

$$\begin{aligned}
\frac{\partial \mathcal{L}(\vec{x},\lambda)}{\partial x_1} &= 1 - \lambda = 0,\\
\frac{\partial \mathcal{L}(\vec{x},\lambda)}{\partial x_2} &= 2 - 2\lambda x_2 = 0,\\
\frac{\partial \mathcal{L}(\vec{x},\lambda)}{\partial x_3} &= 3 - 2\lambda x_3 = 0,\\
\frac{\partial \mathcal{L}(\vec{x},\lambda)}{\partial \lambda} &= 1 - x_1 - x_2^2 - x_3^2 = 0
\end{aligned}$$

Dies führt zum Lösungstupel $(x_1^*, x_2^*, x_3^*, \lambda^*) = (-\frac{9}{4}, 1, \frac{3}{2}, 1)$.

Das Maximierungsproblem der Varianz zur Identifikation der ersten HK lässt sich daher über das Aufspüren des stationären Punktes der folgenden Lagrangiana lösen:

$$\mathcal{L}(\vec{x},\lambda) = \overrightarrow{e_1^T}\boldsymbol{\Sigma}\overrightarrow{e_1} - \lambda[\overrightarrow{e_1^T}\overrightarrow{e_1} - 1] \tag{6}$$

Der stationäre Punkt wird entsprechend wie folgt ermittelt:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\overrightarrow{e_1},\lambda)}{\partial \overrightarrow{e_1}} &= 2\boldsymbol{\Sigma}\overrightarrow{e_1} - 2\lambda\overrightarrow{e_1} = 2(\boldsymbol{\Sigma} - \lambda\boldsymbol{I})\overrightarrow{e_1} = \vec{0} \ \wedge\\
1 - \overrightarrow{e_1^T}\overrightarrow{e_1} &= 0
\end{aligned} \tag{7}$$

Hierbei ist $\boldsymbol{I}$ eine Einheitsmatrix, die bei p Elementen von $\overrightarrow{e_1}$ die Dimensionen $p \times p$ besitzt:

$$\boldsymbol{I} := \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix} \tag{8}$$

Die erste Gleichung in (7) bedeutet, dass die Matrix auf der linken Seite der Gleichung singulär sein muss. Da der Vektor $\overrightarrow{e_1}$ für eine non-triviale Lösung kein Nullvektor sein darf, folgt daraus, dass die Determinante der Matrix $\boldsymbol{\Sigma} - \lambda \boldsymbol{I}$ den Wert null haben muss (vgl. Chatfield und Collins 1980: 59):

$$|\boldsymbol{\Sigma} - \lambda \boldsymbol{I}| = 0 \tag{9}$$

Dies wird am Praxisbeispiel verdeutlicht. Die Kovarianzmatrix der ASFR-Zeitreihen von 30-jährigen Frauen in Deutschland, Italien und Österreich (in dieser Reihenfolge) von 1955-2014 beträgt approximativ:

$$\boldsymbol{\Sigma} \approx \begin{bmatrix} 0{,}00025 & 0{,}00017 & 0{,}00029 \\ 0{,}00017 & 0{,}00039 & 0{,}0003 \\ 0{,}00029 & 0{,}0003 & 0{,}00041 \end{bmatrix}$$

In diesem Fall hat $\boldsymbol{I}$ die Dimensionen 3×3 und Gleichung (9) wird zu:

$$\left| \begin{bmatrix} 0{,}00025 & 0{,}00017 & 0{,}00029 \\ 0{,}00017 & 0{,}00039 & 0{,}0003 \\ 0{,}00029 & 0{,}0003 & 0{,}00041 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right|$$

$$= \left| \begin{bmatrix} 0{,}00025 - \lambda & 0{,}00017 & 0{,}00029 \\ 0{,}00017 & 0{,}00039 - \lambda & 0{,}0003 \\ 0{,}00029 & 0{,}0003 & 0{,}00041 - \lambda \end{bmatrix} \right| = 0$$

Die Determinante einer 3×3-Matrix berechnet sich nach folgendem Schema:

$$\left| \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \right| = aei + bfg + cdh - ceg - bdi - afh$$

Das Schema ist zwar nur bei 3x3-Matrizen gültig, höher dimensionale Matrizen lassen sich jedoch auf einfache Weise in 3x3-Matrizen zerlegen (vgl. hierzu z.B. Simon und Blume 1994: 190-193), worauf an dieser Stelle aber nicht weiter eingegangen werden soll. Das bedeutet im Beispiel, dass sich das Lösungsproblem vereinfacht zu:

$$(0{,}00025 - \lambda)(0{,}00039 - \lambda)(0{,}00041 - \lambda) + 0{,}00017 \cdot 0{,}0003 \cdot 0{,}00029 +$$
$$0{,}00029 \cdot 0{,}00017 \cdot 0{,}0003 - 0{,}00029 \cdot (0{,}00039 - \lambda) \cdot 0{,}00029$$
$$-0{,}00017 \cdot 0{,}00017 \cdot (0{,}00041 - \lambda) - (0{,}00025 - \lambda) \cdot 0{,}0003 \cdot 0{,}0003 = 0$$

Die Lösungen der Gleichung sind $\lambda_1 \approx 0{,}00088, \lambda_2 \approx 0{,}00016$ und $\lambda_3 \approx 0{,}00001$. Dies sind die Eigenwerte (EW) der Kovarianzmatrix. Es lässt sich zeigen, dass die EW in ihrer Summe der Kovarianz der Originalvariablen entsprechen. Deshalb werden sie absteigend angeordnet. Wie bereits zuvor erklärt, ist einer der beiden Gründe für die Hauptkomponentenanalyse die Reduktion der

ursprünglichen Problematik auf wenige Variablen, die einen möglichst hohen Anteil der Kovarianz der Originalvariablen erklären. Das bedeutet, dass der erste EW der Varianz entspricht, die von der ersten HK erklärt wird. Allgemein folgt daraus, dass für $i < p$ der folgende Quotient den Anteil der Gesamtvarianz der p Variablen angibt, die von den ersten i HK erklärt werden (vgl. Chatfield und Collins 1980: 58-61):

$$Q := \frac{\sum_{j=1}^{i} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \tag{10}$$

Im Beispiel beträgt $Q \approx 0{,}839$ für i=1 und 0,988 für i=2. Das bedeutet, die beiden ersten HK würden in diesem Fall bereits ausreichen, um fast 99% der Variation in den ASFR der 30-jährigen in Deutschland, Italien und Österreich zu erklären.

Der nächste wichtige Schritt ist die Bestimmung der zugehörigen Eigenvektoren (EV). Diese werden sequenziell durch Einsetzen der EW in (7) ermittelt (vgl. Handl: 123-124). Für λ_1 bedeutet das:

$$\begin{bmatrix} 0{,}00025 - 0{,}00088 & 0{,}00017 & 0{,}00029 \\ 0{,}00017 & 0{,}00039 - 0{,}00088 & 0{,}0003 \\ 0{,}00029 & 0{,}0003 & 0{,}00041 - 0{,}00088 \end{bmatrix} \overrightarrow{e_1} = \vec{0}$$

Dies stellt ein Lineares Gleichungssystem dritten Grades dar

$$\begin{bmatrix} -0{,}00063e_{11} + 0{,}00017e_{12} + 0{,}00029e_{13} \\ 0{,}00017e_{11} - 0{,}00049e_{12} + 0{,}0003e_{13} \\ 0{,}00029e_{11} + 0{,}0003e_{12} - 0{,}00047e_{13} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

, wobei z.B. e_{11} das erste Element des ersten EV ist. Die Lösung des Gleichungssystems ergibt den ersten EV:

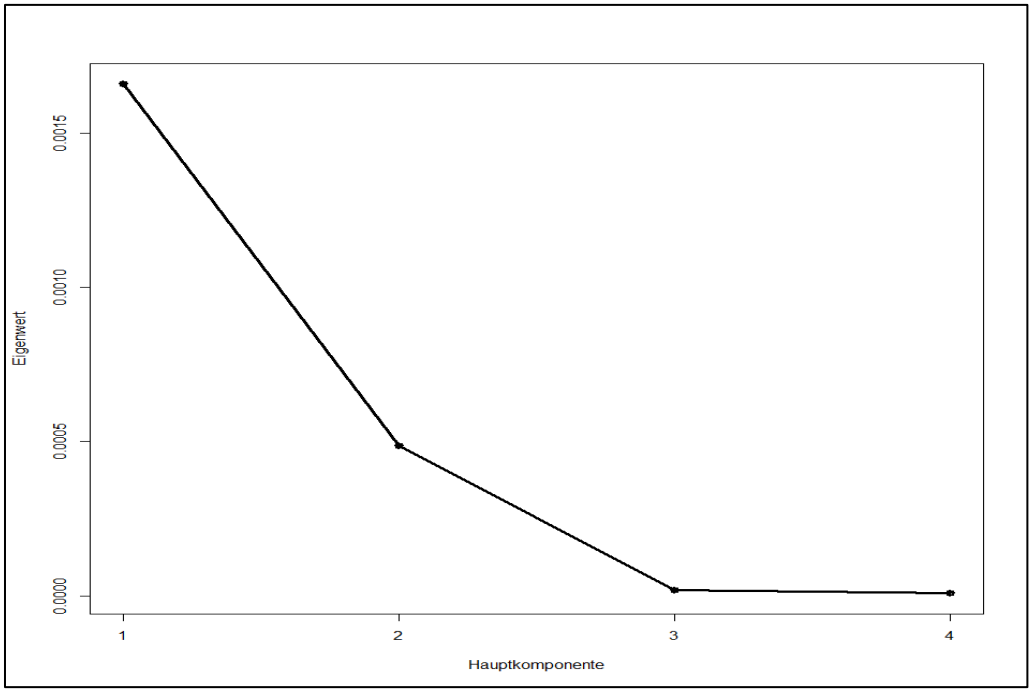
$$\overrightarrow{e_1} = \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \end{bmatrix} \approx \begin{bmatrix} -0{,}47012 \\ -0{,}57997 \\ -0{,}6653 \end{bmatrix}$$

Dabei können die Elemente des EV als Korrelationen mit den Originalvariablen angesehen werden. Die erste HK ist entsprechend sowohl mit den ASFR der 30-jährigen in Deutschland, Italien und Österreich relativ stark negativ korreliert, am stärksten mit Österreich. $\overrightarrow{e_2}$ und $\overrightarrow{e_3}$ lassen sich auf gleiche Weise ermitteln, worauf an dieser Stelle nicht weiter eingegangen werden soll. Final lassen sich an dieser Stelle wie in (1) beschrieben die HK selbst kalkulieren.

Eine wichtige Frage ist die Bestimmung der Anzahl an HK, die für weitere Analysen genutzt werden. Es gibt auf diese Frage keine triviale Antwort. Modellierer müssen subjektiv entscheiden, wie viele HK sie nutzen. Es gibt hierzu jedoch eine Reihe von Kriterien, die die Entscheidung vereinfachen können. Eine Möglichkeit liegt darin, einen Grenzwert an Gesamtstreuung vorzugeben, der von den HK erklärt werden soll. Um die Konzepte besser verdeutlichen zu können, wird an dieser Stelle zusätzlich zu den drei vorigen Variablen noch die ASFR der 30-jährigen in Spanien hinzugezogen. In diesem Fall wären die vier EW $\lambda_1 \approx 0{,}00166, \lambda_2 \approx 0{,}00049$, $\lambda_3 \approx 0{,}00002$ und $\lambda_4 \approx 0{,}00001$. Es werden dann so viele HK gewählt, dass *Q* in (10) diesen Grenzwert gerade übertrifft. Ist das Ziel bspw., so viele HK zu verwenden, dass mindestens 90% der Gesamtstreuung erklärt werden, so sind die ersten zwei HK zu wählen, da diese rund 98,8% erklären, während die erste HK allein 76,4% erklärt.

Eine weitere gängige Methode ist das Heranziehen eines Screeplots. Dabei werden die Eigenwerte der Kovarianzmatrix grafisch abgetragen, wie in Abbildung 1 zu sehen.

Abbildung 1: Screeplot der Hauptkomponenten



Quelle: Eigene Berechnung und Darstellung

Nach diesem Kriterium sollten lediglich die Hauptkomponenten ins Modell einbezogen werden, die sich vor dem Knick befinden, wobei es keine eindeutige Definition gibt, ob die Hauptkomponente „im Knick“ selbst auch inkludiert werden sollte. Der Screeplot würde implizieren, ein bis

zwei HK zu nutzen. Aus praktischer Sicht lohnt es sich, für Prognosen in der Regel, die HK im Knick einzubeziehen, da ansonsten häufig ein relativ großer Anteil der Varianz ignoriert wird, was im Hinblick auf die Konstruktion von Prognoseintervallen zu Verklärungen führen könnte. Dies ist ein in der praktischen Anwendung häufig beobachtbarer Fehler. Vanella et al. (2017) schlagen eine Methode vor, den aus dem Weglassen einiger HK entstehenden Schätzfehler in der zukünftigen Unsicherheit einzubeziehen, um zu schmale Prognoseintervalle zu verhindern. Darauf wird im Rahmen dieser Ausarbeitung jedoch nicht weiter eingegangen.

Andere Alternativen sind Kaisers Kriterium und das Kriterium von Jolliffe. Kaisers Kriterium besagt, dass alle HK genutzt werden sollten, deren zugehörige Eigenwerte über dem Mittelwert der Eigenwerte liegen. Im Beispiel liegt der Mittelwert der Eigenwerte bei etwa 0,00054. Der einzige EW der über diesem Wert liegt, ist entsprechend der erste, weshalb nur die erste HK gewählt würde. Jolliffe riet dazu, stattdessen lediglich das 0,7fache des Mittelwertes der EW als Grenze zu wählen, was im Beispiel etwa 0,00038 ist. Nach Jolliffe würden daher die ersten zwei HK inkludiert (vgl. Handl 2010: 128-129).

3 Grundzüge der Zeitreihenanalyse

In diesem Abschnitt wird auf einige wenige Aspekte der Zeitreihenanalyse eingegangen, die für die Hauptkomponentenprognose von hoher Relevanz sind.

Eine Zeitreihe ist eine Variable, die in jeder Periode eine Beobachtung generiert. Das fundamentale Konzept der modernen Zeitreihenanalyse ist die Stationarität, weshalb diese an dieser Stelle kurz erklärt werden muss. Die Zeitreihe der ASFR (in Periode t) der 30-jährigen Frauen in Deutschland werde an dieser Stelle mit a_t bezeichnet. Für die Stationarität von a_t sind zwei Bedingungen hinreichend, die Erwartungswertstationarität und die Autokovarianzstationarität (vgl. Shumway und Stoffer 2011: 23).

Erwartungswertstationarität bedeutet, dass der Erwartungswert der Zeitreihe in jeder Periode gleich ist:

$$E[a_s] = E[a_t] \; \forall \; s, t \tag{11}$$

Autokovarianzstationarität besagt, dass die Autokovarianz zwischen zwei Beobachtungen der Zeitreihe unabhängig von der Zeit ist und lediglich davon abhängt, wie lange das Zeitintervall zwischen den beiden Beobachtungen ist:

$$Cov[a_{t+h}, a_t] = Cov[a_h, a_0] \; \forall \; t \tag{12}$$

Für die praktische Anwendung von besonderer Bedeutung ist das von Box und Jenkins definierte Autoregressive Integrated Moving Average (ARIMA)-Modell. Dieses wird nachfolgend sequenziell erklärt (vgl Shumway und Stoffer 2011: 84-93). Ein MA(q)-Modell ist wie folgt definiert:

$$a_t = \omega_t - \sum_{i=1}^{q} \theta_i \omega_{t-i} \tag{13}$$

Dabei ist ω_t ein stochastischer Störterm in Periode t, welcher aus praktischer Sicht häufig als normalverteilt angenommen wird mit Erwartungswert null und einer Varianz σ^2:

$$\omega_t \sim \mathcal{NID}(0, \sigma^2) \tag{14}$$

An dieser Stelle ist die Stationaritätsannahme von hoher Bedeutung. Stationarität bedeutet, dass der Störterm als identisch verteilt in jeder Periode angenommen werden kann. Diese Annahme ist vor allem für die Durchführung von Simulationen empfehlenswert. Das MA(q)-Modell geht folglich davon aus, dass sich die aktuelle Beobachtung der Variablen ausschließlich als gewichtete Summe der q letzten Ausprägungen des Störterms und dem Störterm der aktuellen Periode ergibt. Dabei ist θ_i der Korrelationskoeffizient der Zeitreihe bzgl. des Fehlers in Periode t-i. θ_i darf dabei nur Werte zwischen -1 und 1 einnehmen:

$$|\theta_i| < 1 \tag{15}$$

Eine praktikable alternative Schreibweise eines MA(q) ist die sogenannte Lag-Schreibweise, wobei *L* der sogenannte *Lag-Operator*[3] ist. Die Lag-Schreibweise für das MA(q)-Modell ist:

$$a_t = \left(1 - \sum_{i=1}^{q} \theta_i L^i\right) \omega_t \tag{16}$$

Der Exponent von *L* gibt dabei an, auf welche vergangene Periode sich bezogen wird. $L^q \cdot \omega_t$ bedeutet z.B. ω_{t-q}.[4]

[3] Alternativ wird in der Literatur auch vom *Backshift-Operator* gesprochen.

[4] In der praktischen Anwendung muss auf die genaue Definition der Koeffizienten geachtet werden. Manche Statistik-Programme geben leicht unterschiedliche Ergebnisse aus. So sind bspw. beim ***R***-Output die Vorzeichen der Koeffizienten im Vergleich zur Definition in diesem Beitrag umgekehrt.

Ein weiterer einfacher Fall eines Zeitreihenmodells ist ein AR(p):

$$a_t = \omega_t + \sum_{j=1}^{p} \phi_j a_{t-j} \tag{17}$$

bzw. in Lag-Schreibweise:

$$\left(1 - \sum_{i=1}^{p} \phi_j L^j\right) a_t = \omega_t \tag{18}$$

Im AR(p)-Modell wird entsprechend die Zeitreihe in t auf ihre letzten p Beobachtungen regressiert (wobei auch hier der Störterm) der Periode t einbezogen wird. Analog bedeutet diesem Fall $L^p \cdot a_t = a_{t-p}$. Auch hier gilt für die Korrelationskoeffizienten:

$$|\phi_j| < 1 \tag{19}$$

AR- und MA-Modelle lassen sich auch kombinieren, sodass die Kombination aus einem AR(p) und einem MA(q)-Modell zu einem ARMA(p,q)-Modell führt, welches formal wie folgt definiert ist:

$$a_t = \omega_t - \sum_{i=1}^{q} \theta_i \omega_{t-i} + \sum_{j=1}^{p} \phi_j a_{t-j} \tag{20}$$

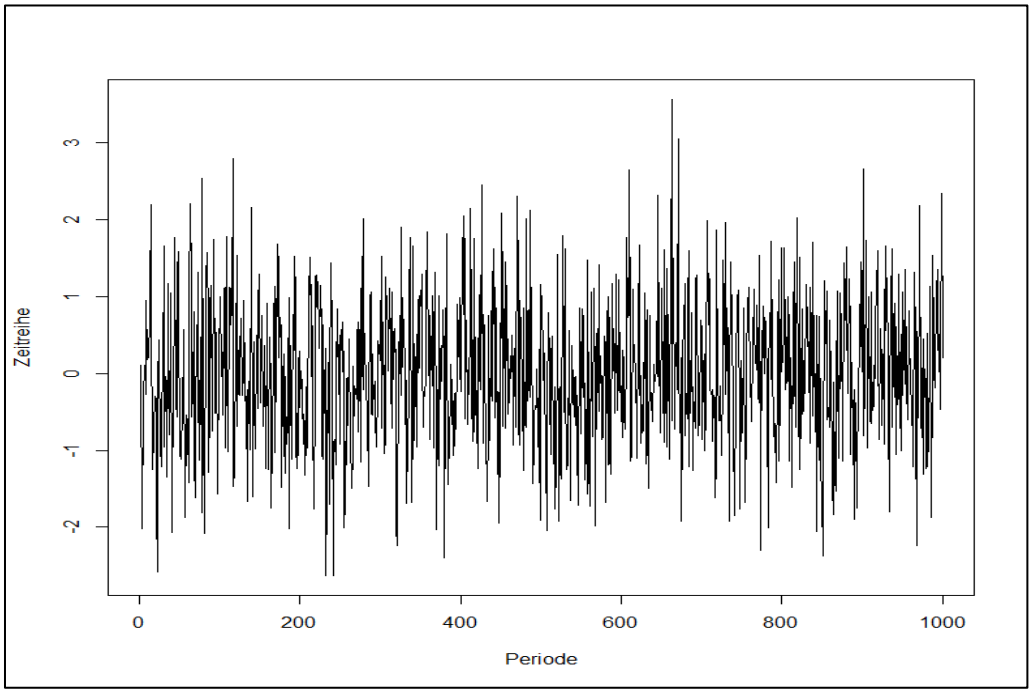
bzw.

$$\left(1 - \sum_{i=1}^{p} \phi_j L^j\right) a_t = \left(1 - \sum_{i=1}^{q} \theta_i L^i\right) \omega_t \tag{21}$$

Wie bereits angedeutet, ist bei ARMA-Prozessen die Stationaritätsannahme fundamental. Es stellt sich die Frage, woran ein Forscher also erkennt, ob seine Zeitreihe stationär ist. Dafür empfiehlt sich im ersten Schritt die grafische Darstellung der Zeitreihe. Abbildung 2 zeigt eine simulierte stationäre Zeitreihe[5]. Es lässt sich beobachten, dass weder der Erwartungswert, noch die Varianz der Zeitreihe einem Trend unterliegen.

[5] Die Zeitreihe wurde durch 1000-fache Computersimulation einer standardnormalverteilten Zufallsvariable erzeugt.

Abbildung 2: Stationäre Zeitreihe



Quelle: Eigene Berechnung und Darstellung

Weiterhin sollte die Stationaritätshypothese auf Basis statistischer Tests überprüft werden. Gängige Tests sind dabei u.a. der Augmented Dickey-Fuller *(ADF)*-Test und der Kwiatkowski-Phillips-Schmidt-Shin *(KPSS)* Test. Der ADF-Test testet im einfachsten Fall dabei die Nullhypothese, dass der Prozess ein Random Walk ist, also das für die nachfolgende Gleichung

$$a_t = \rho \cdot a_{t-1} \tag{22}$$

gilt

$$H_0: \rho = 1,$$

was einem Random Walk Prozess entspricht (vgl. Dickey und Fuller 1979: 427). Es gibt unterschiedliche Varianten für den Test. Hier interessiert diejenige mit der Alternativen

$$H_1: |\rho| < 1,$$

was einem stationären bzw. asymptotisch stationären Prozess entspricht. Die Teststatistik ist in diesem Fall

$$\tau = \frac{\hat{\rho} - 1}{se(\hat{\rho})} \tag{23}$$

, was einem herkömmlichen t-Test entspricht. Dies wird allerdings nicht mit den Quantilen einer t-Verteilung, sondern einer empirischen Verteilung in Kontrast gesetzt, die von Dickey und Fuller

auf Basis von Monte Carlo-Simulationen kalkuliert wurde (vgl. Fuller 1996: 642).[6] Im Beispiel liegt τ bei etwa -9.8868, was heißt, dass die Nullhypothese auf allen gängigen Konfidenzniveaus verworfen würde. Dies ließe sich so interpretieren, dass die statistische Evidenz darauf hindeuten würde, dass die Zeitreihe stationär wäre.

Der KPSS-Test hingegen ist ein Lagrange-Multiplier-Test mit Test-Statistik

$$LM = \frac{\sum_{t=1}^{T} S_t^2}{SSR/T} \tag{24}$$

, wobei SSR für die Summe der quadrierten Residuen der Regression steht, T die Anzahl der Perioden ist und S_t die Summe der Residuen aus der Regression

$$a_t = \alpha + \beta t$$

bis zum Zeitpunkt t darstellt. Dabei wurden die kritischen Werte der zugrundeliegenden Verteilung von Kwiatkowski et al. über einen Wiener-Prozess[7] generiert, worauf an dieser Stelle nicht weiter eingegangen werden soll.[8] Der KPSS-Test testet dabei die Nullhypothese, dass die fragliche Zeitreihe stationär ist. Entsprechend ergeben sehr große Werte ein Verwerfen von H_0 (vgl. Kwiatkowski et al. 1992: 162-167). Im Beispiel ergibt sich für die Teststatistik ein Wert von etwa 0,2222 für den Standard-Test. Damit liegt dieser unter den kritischen Werten auf allen gängigen Niveaus[9] und die Nullhypothese kann nicht verworfen werden. Der KPSS-Test würde dementsprechend keine Evidenz liefern, die Annahme der Stationarität zu verwerfen. Schließlich ist noch darauf zu testen, ob die Fehler im Modell tatsächlich homoskedastisch sind, also die Fehler in jeder Periode die gleiche Varianz aufweisen. Dies kann anhand verschiedener Tests geschehen. An dieser Stelle sei beispielhaft Engles ARCH-LM Test genannt. Dieser ist ein vergleichsweise einfacher LM-Test, der die Nullhypothese testet, dass die Varianz, unabhängig von t, über alle Perioden konstant ist.

[6] Der ADF-Test ist standardmäßig in Statistik-Programmen implementiert, z.B. in ***R*** unter dem Befehl **adf.test**, der im Paket **tseries** enthalten ist.

[7] Dabei wächst der Prozess in jeder Periode um einen stochastischen Wert, der aus einer stationären normalverteilten Zufallsvariable gezogen wird.

[8] Der KPSS-Test ist in jeder guten Statistik-Software standardmäßig integriert. In ***R*** z.B. lässt er sich mit dem Befehl **kpss.test** aus dem Paket **tseries** durchführen.

[9] Für α=0,1 liegt der kritische Wert z.B. bei rund 0,347.

Es würde den Rahmen dieser Arbeit sprengen, den Test im Detail zu erklären, daher sei an interessierte Leser einfach der Hinweis auf die zugehörige Arbeit von Engle gerichtet (Vgl. Engle 1982).[10]

Ist der Modellierer aufgrund der Testergebnisse der Meinung, die Zeitreihe sei noch nicht stationär, so wird eine Transformation benötigt. Es wird in diesem Fall angenommen, die Zeitreihe sei integriert, was den mittleren Teil des ARIMA-Ausdrucks betrifft. Eine Zeitreihe, die d-fach integriert ist, wird im einfachsten Falle als ARIMA(0,d,0)-Prozess bezeichnet (vgl. Shumway und Stoffer 2011: 141):

$$(1 - L)^d a_t = \omega_t \tag{25}$$

Eine nicht stationäre Zeitreihe lässt sich grds. durch eine gewisse Anzahl von Differenzenbildungen in eine stationäre umwandeln (vgl. Shumway und Stoffer: 58-61). Dabei wird die erste Differenz der Zeitreihe wie folgt gebildet:

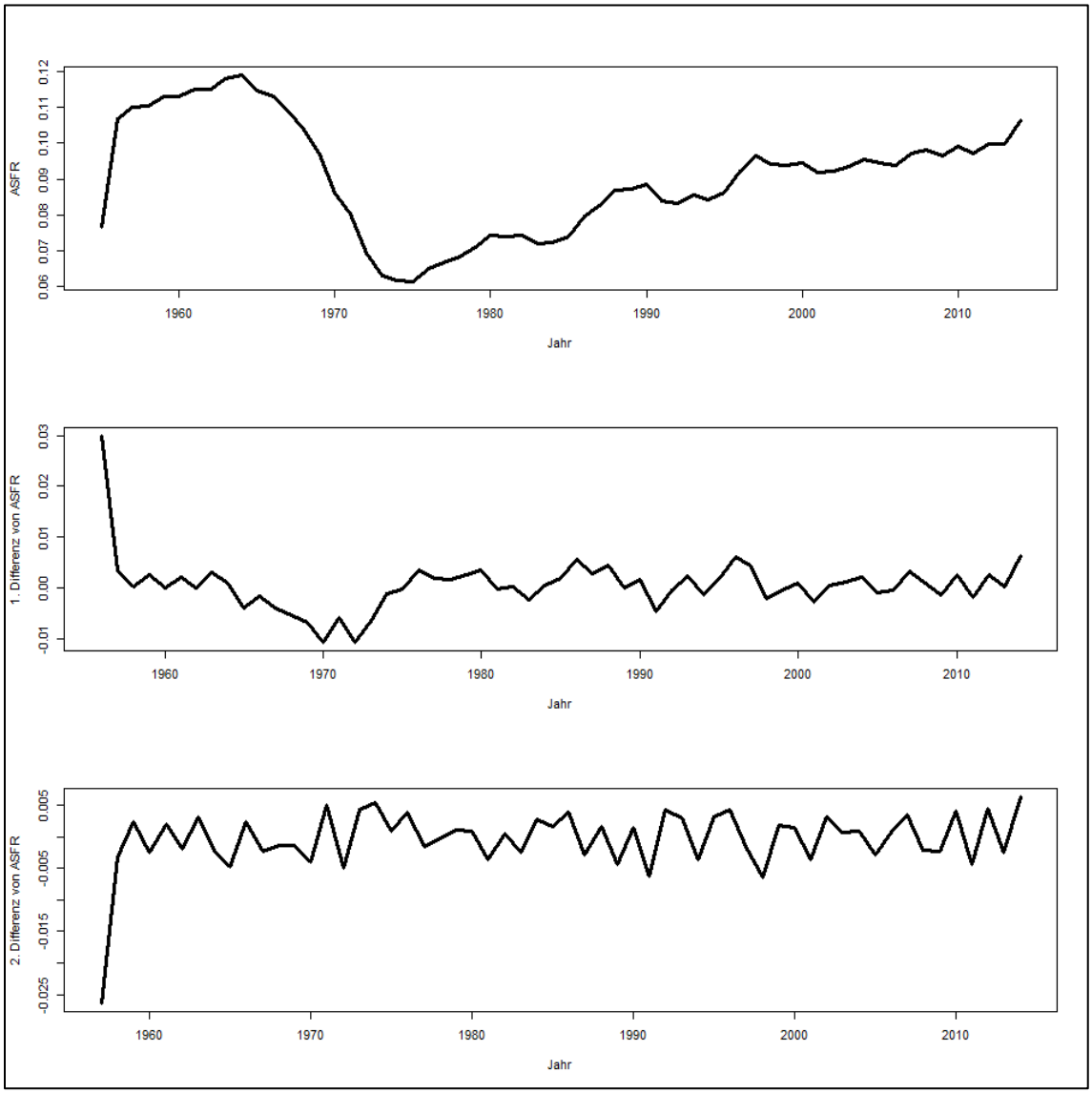
$$\Delta a_t = a_t - a_{t-1} \tag{26}$$

Wie aus der Analysis bekannt, sorgt diese Operation approximativ dafür, dass die Potenz der Zielfunktion, hier der Zeitreihe, um eins verringert wird.

Als Beispiel für die Folgen der Differenzenbildung zeigt Abbildung 3 die Zeitreihe der ASFR der 30-jährigen in Deutschland mit ihrer ersten und zweiten Differenz. Es ist zu beobachten, dass die Trends der ursprünglichen Zeitreihe deutlich flacher werden und die zweite Differenz grafisch wie eine stationäre Zeitreihe aussieht. Ist die Zeitreihe zu einer vermeintlich stationären umgewandelt, so stellt sich bei dieser schließlich die Frage, welches ARMA-Modell die (asymptotisch) stationäre Zeitreihe am besten schätzt.

[10] Auch der ARCH-LM Test ist standardmäßig in Statistik-Programmen integriert. In **R** lässt er sich über den Befehl **ArchTest** im **FinTS**-Paket durchführen.

Abbildung 3: ASFR mit ersten und zweiten Differenzen

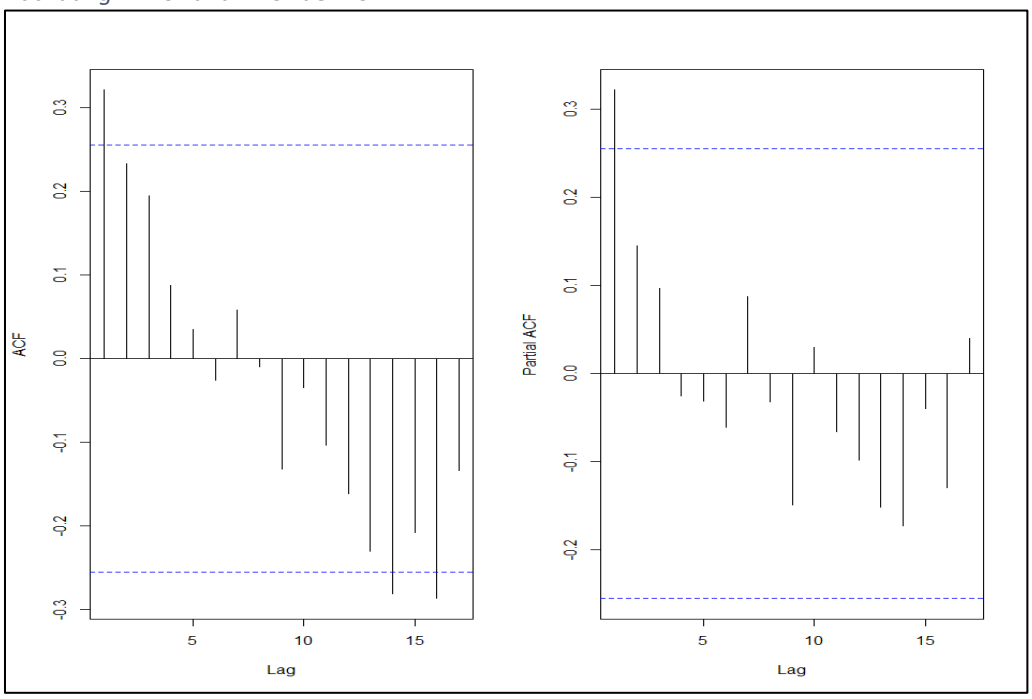


Quelle: Eigene Berechnung und Darstellung

Dazu gibt es eine Reihe von Informationskriterien, die herangezogen werden können, vorrangig das Akaike Information Criterion (*AIC)*, das Bayesian Information Criterion *(BIC)*, welches auch als Schwartz Kriterium bekannt ist (siehe hierzu z.B. Greene 2012: 180), und das Hannan-Quinn-Criterion *(HQC) (siehe Hannan und Quinn 1979: 191)*. Diese sind in ihrer Form verwandt und beziehen sich auf die Log-Likelihood der Anpassung als Gütekriterium. Der Unterschied liegt dabei an dem Ausmaß, in dem diese die Komplexität des Modells bestrafen. Ziel dabei ist die Wahl des Modells, welches das ausgewählte Kriterium minimiert. Auf die Kriterien soll an dieser Stelle nicht weiter eingegangen werden, da sie sehr stark von Asymptotik abhängen. Die Anpassungsgüte der Informationskriterien ist daher sehr abhängig davon, wie lange die Zeitreihe ist (allgemein: wie

viele Daten zur Verfügung stehen), die als Input ins Modell gesteckt wird. Da die Datenverfügbarkeit und Qualität in typischen bevölkerungswissenschaftlichen Fragestellungen, besonders in Bezug auf Bevölkerungsprognosen, relativ gering ist, sind die Informationskriterien also eher kritisch zu betrachten, obwohl sie durchaus ergänzend genutzt werden können. Für Zeitreihen der Fertilität oder Mortalität empfehlen sich in erster Linie grafische Analysen in Form der Autokorrelationsfunktion *(ACF)* und der Partiellen Autokorrelationsfunktion *(PACF)*[11]. Für das Beispiel der ASFR empfiehlt der ADF-Test zweimaliges Differenzieren, während der KPSS-Test eher zu einer Differentiation rät. Auch in diesem Fall ist allerdings praktisch zu erwähnen, dass der KPSS-Test bei geringer Historie eher schwach abschneidet und tendenziell zu früh Stationarität anzeigt. In Verbindung mit Abbildung 3 könnte davon ausgegangen werden, dass bereits eine Differenzenbildung zu einer stationären Zeitreihe führt, da der ADF-Test, wie bei allen statistischen Tests üblich, sehr abhängig von einer großen Datenmenge, die hier nicht verfügbar ist. Abbildung 4 illustriert für das Beispiel die ACF und PACF.

Abbildung 4: ACF und PACF der ASFR



Quelle: Eigene Berechnung und Darstellung

Die Grafiken geben einen Hinweis darauf, welche Lag-Länge gewählt werden sollte, also die Werte, die für p und q gewählt werden sollten. Die grafische Analyse ist nicht trivial und erfordert

[11] Für eine detaillierte Beschreibung von ACF und PACF siehe z.B. Shumway und Stoffer 2011: 102-108.

ein gewisses Maß an Erfahrung durch den Anwender. Es gibt allerdings ein paar Eigenschaften von AR und MA-Prozessen, die im Idealfall an den Grafiken zu erkennen sein könnten. Zum einen gibt die gestrichelte Linie[12] Hinweise auf die Lag-Länge. Ein AR(1)-Prozess lässt sich meistens relativ leicht erkennen, da bei diesem zumeist die ACF exponentiell abfällt, während die PACF beim ersten Lag hoch ist und danach direkt auf etwa null abfällt. Der MA(1) verhält sich genau umgekehrt. In Abbildung 4 lässt sich beobachten, dass die ACF relativ monoton abfällt, während die PACF bereits nach dem ersten Lag deutlich sinkt. Daher lassen die Grafiken vermuten, dass die erste Differenz der ASFR-Zeitreihe am besten durch einen AR(1)-Prozess charakterisiert werden kann. Die ASFR ist in diesem Fall folglich ein ARIMA(1,1,0)-Prozess. Diese Vermutung lässt sich anhand der Informationskriterien prüfen. In diesem Fall führt die Minimierung des AIC tatsächlich zu einem ARIMA(1,1,0).[13]

An dieser Stelle wird darauf eingegangen, wo der Nutzen der oben angesprochenen Lag-Schreibweise liegt[14]. Beim allgemeinen ARIMA(p,d,q)-Prozess stellt sich diese wie folgt dar:

$$\left(1 - \sum_{i=1}^{p} \phi_j L^j\right)(1-L)^d a_t = \left(1 - \sum_{i=1}^{q} \theta_i L^i\right)\omega_t \tag{27}$$

Im Fall eines ARIMA(1,1,1) ist die Lag-Schreibweise:

$$(1-\phi L)(1-L)a_t = (1-\theta L)\omega_t$$

Dies lässt sich ausmultiplizieren zu:

$$[1-(1+\phi)L+\phi L^2]a_t = (1-\theta L)\omega_t$$

Aus der Definition des Lag-Operators folgt somit:

$$a_t - (1+\phi)a_{t-1} + \phi a_{t-2} = \omega_t - \theta\omega_{t-1}$$

Und damit schließlich:

$$a_t = (1+\phi)a_{t-1} - \phi a_{t-2} + \omega_t - \theta\omega_{t-1}$$

[12] Die Grafiken wurden in ***R*** mit dem **acf()** und dem **pacf()**-Befehl erzeugt. Die gestrichelten Linien gibt das Programm automatisch mit aus, je nach dem Signifikanzniveau, welches gewählt wird.

[13] Es gibt in der Software Standard-Algorithmen, die die Optimierung durchführen. Bei ***R*** wurde dies mit dem **auto.arima**-Befehl aus dem **forecast**-Paket geprüft, der in der Tat einen ARIMA(1,1,0) angezeigt hat.

[14] Da in dieser Arbeit nicht zu viele statistische Details eingebracht werden sollen, wird nicht auf den weiteren Vorteil der leichteren Umformung in die Lineare Zeitreihenschreibweise eingegangen.

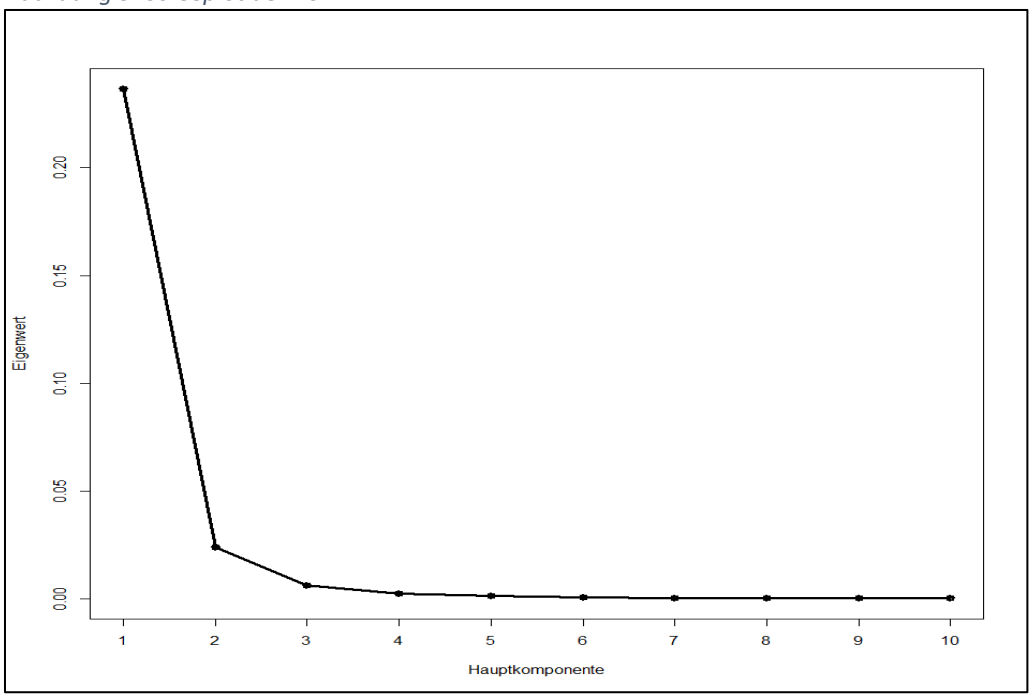
Mit Hilfe des Lag-Operators lassen sich also auch relativ schwierig wirkende funktionale Formen in handhabbarer Art schreiben, was in der Prognose im Rahmen von Simulationsstudien besonders hilfreich ist.

4 Prognose altersspezifischer Fertilitäts- und Überlebensraten

In diesem Abschnitt wird vorgestellt, wie die Methoden der Zeitreihenanalyse dazu genutzt werden können, um die zuvor identifizierten Hauptkomponenten zu prognostizieren. Einen ersten vergleichbaren Ansatz hierzu schlugen Bell und Monsell für die Prognose altersgruppenspezifischer Mortalität in den Vereinigten Staaten vor (vgl. Bell und Monsell 1991). Die Abwandlung dieser Arbeit durch Lee und Carter für Prognosen der nach Altersklassen aufgeteilten Mortalitäts- und Fertilitätsraten in der Vereinigten Staaten ist auch heute noch sehr populär (siehe hierzu Carter und Lee 1992 und Lee 1993). Einige deutsche Wissenschaftler benutzen bereits seit einigen Jahren das Lee-Carter-Modell zur Prognose alters- und geschlechtsspezifischer Fertilitäts- und Mortalitätsraten (vgl. z.B. Härdle und Myšičkova 2005: 5-14; Deschermeier 2015). Der Autor des vorliegenden Beitrages nutzt an dieser Stelle einen ähnlichen, aber in Bezug auf die Datengrundlage und die ARIMA-Modelle abweichenden Ansatz. Dazu wurden für 16 europäische Länder die ASFR für die Jahre 1973-2014 aus der Human Fertility Database, der Human Fertility Collection, Eurostat sowie der obersten statistischen Behörden in der Schweiz und Portugal gesammelt. Es wurden dabei die ASFR für 13- bis 54jährige extrahiert, was in einer Gesamtanzahl von 672 Variablen resultiert. Auf Basis dieser Daten sollen die ASFR für Frauen in Deutschland prognostiziert werden. Um auch internationale Trends abbilden zu können, nutzt der Autor die internationalen Daten zur Prognose der deutschen Fertilitätsraten. Das hat den Vorteil, dass der relativ geringen Datenmenge entgegengewirkt wird. Die Hauptkomponentenanalyse bietet die Möglichkeit, die Korrelationen zwischen den ASFR zu nutzen, ohne dabei eine Verzerrung zu erzeugen, da, wie in Abschnitt 2 erwähnt, bei nicht vorhandenen Korrelationen die Originalvariablen wieder ausgegeben werden. An dieser Stelle ist es offensichtlich, dass eine individuelle Prognose von 672 Variablen nicht nur extrem aufwändig und damit nahezu unmöglich regelmäßig durchzuführen wäre, es würden dabei auch die Korrelationen zwischen den altersspezifischen Fertilitätstrends ignoriert.

Beide Probleme werden von der Hauptkomponentenanalyse angegangen. Die Hauptkomponentenanalyse zeigt dabei, dass bereits die ersten zwei HK etwa 95,3% der Variation in allen 672 ASFR erklären. Über 97,5% werden von den ersten drei HK erklärt. Abbildung 5 zeigt den Screeplot für die ASFR. Würde der Modellierer festlegen, 95% erklärter Varianz seien hinreichend und orientiert sich am Screeplot, der zu zwei HK rät, so würden zwei HK reichen, um eine adäquate Prognose aller ASFR durchzuführen. Der Autor wählt aus Gründen größerer Genauigkeit alle HK aus, die mindestens 1% der Gesamtstreuung erklären, was in diesem Fall die ersten drei HK betrifft.

Abbildung 5: Screeplot der ASFR



Quelle: Eigene Berechnung und Darstellung

Die Korrelationen der ersten drei HK mit den ASFR in Deutschland werden in Abbildung 6 dargestellt.

Nun wird am Beispiel der ersten HK die Prognose erläutert. Wie in Abschnitt 3 beschrieben, wurde ein ARIMA-Modell an den historischen Verlauf der HK angepasst. Dabei stellt sich das folgende ARIMA(1,2,0)-Modell als am besten geeignet heraus:

$$(1 + 0{,}54186L)(1 - L)^2 p_t = \varepsilon_t$$

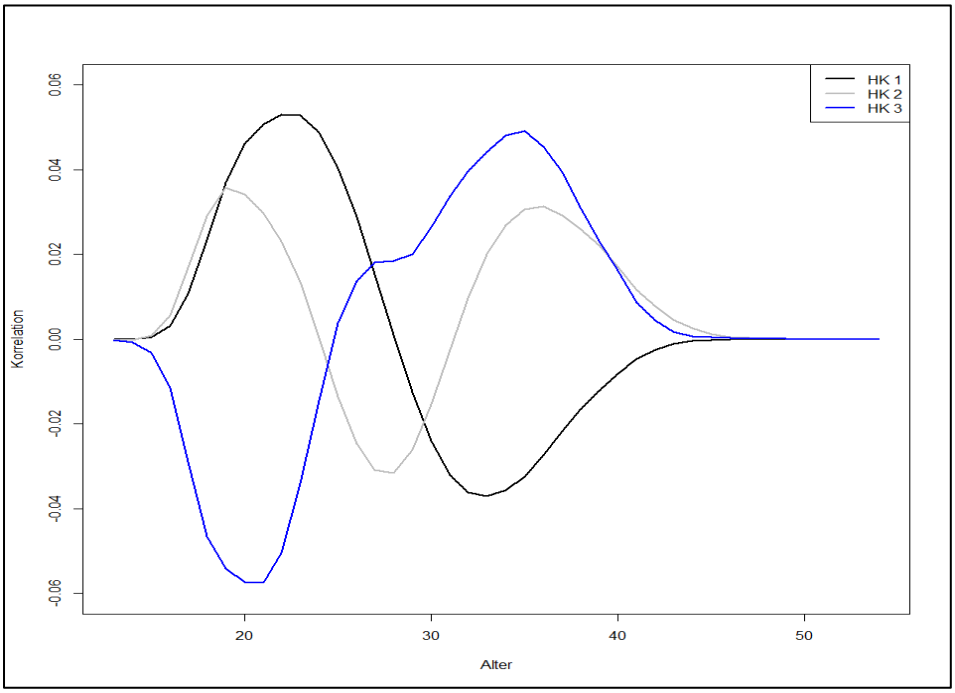
Dies vereinfacht sich nach Ausmultiplikation zu:

$$p_t = 1{,}45814 p_{t-1} + 0{,}08372 p_{t-2} - 0{,}54186 p_{t-3} + \varepsilon_t \qquad (28)$$

Da der Erwartungswert des Störterms bei null liegt, folgt daraus:

$$E[p_t] = 1{,}45814p_{t-1} + 0{,}08372p_{t-2} - 0{,}54186p_{t-3} \quad (29)$$

Abbildung 6: Korrelationen zwischen ASFR und HK

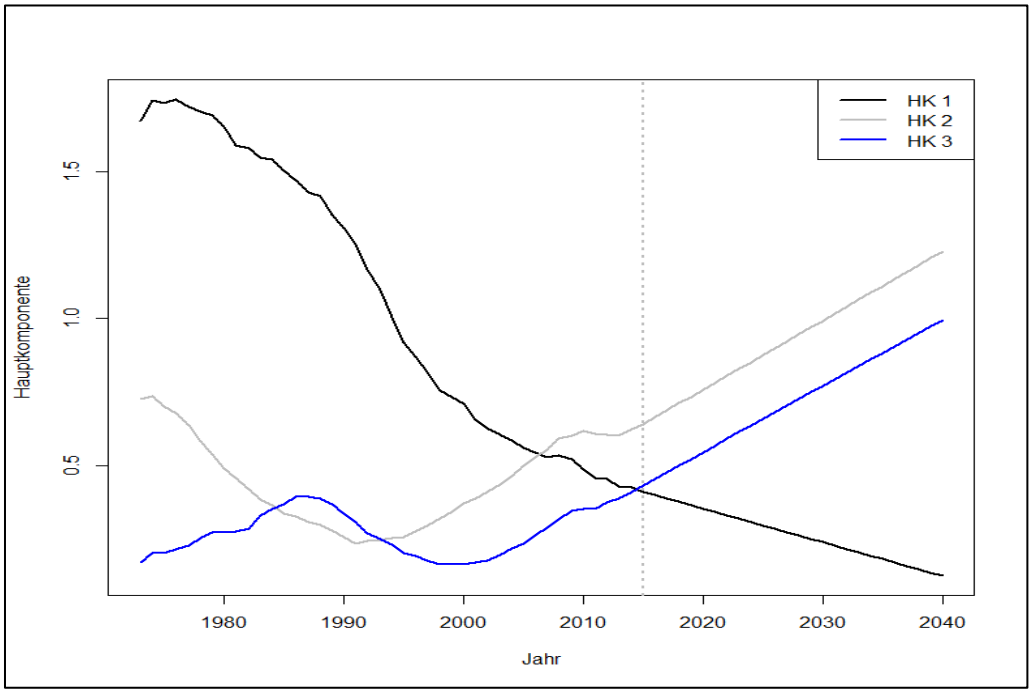


Quelle: Eigene Berechnung und Darstellung

Auf Basis dieser Anpassung lässt sich der erwartete Pfad der ersten HK bis ins Jahr 2040 prognostizieren. Auf die gleiche Weise lässt sich dies für die beiden weiteren HK durchführen, deren funktionale Form an dieser Stelle nicht erwähnt wird, da die Erklärung der Methodik im Fokus steht. Im konkreten Fall ergeben sich als wahrscheinlichste Szenarien für den Verlauf der ersten drei HK bis 2040 die folgenden Kurven:

Abbildung 7 zeigt den historischen Verlauf der ersten drei HK seit 1973. Die graue gestrichelte Linie markiert das erste Prognosejahr 2015. Die verbleibenden 669 Hauptkomponenten werden für die Prognose konstant auf dem Niveau von 2014 angenommen. Da diese lediglich knapp 2,5% der Gesamtstreuung der ASFR für alle 16 Länder erklären, ist der Fehler, der durch diese Annahme entsteht, verschwindend klein, während die Modellroutine gleichzeitig stark vereinfacht wird.

Abbildung 7:Historischer Verlauf und Prognose der Hauptkomponenten



Quelle: Eigene Berechnung und Darstellung

Die Historie der Hauptkomponenten wird dabei durch die Verallgemeinerung von Gleichung (1) erzeugt:

$$\boldsymbol{P} = \boldsymbol{FE} \tag{30}$$

Dabei ist ***P*** eine Matrix mit t Zeilen und s Spalten, wobei t in diesem Fall die Anzahl der beobachteten Perioden und s die Anzahl der Zeitreihen ist. Im Beispiel hat P folglich die Dimensionen 42×672. ***F*** ist die Matrix aller ASFR und genauso strukturiert wie ***P***, also eine spaltenweise Zusammenstellung der Zeitreihen aller ASFR. Dabei ist die Reihenfolge der Zeitreihen irrelevant für die Hauptkomponentenanalyse, da diese die HK sowieso nach ihrer Relevanz ordnet. ***E*** ist eine Matrix, die die Eigenvektoren spaltenweise aneinanderreiht.

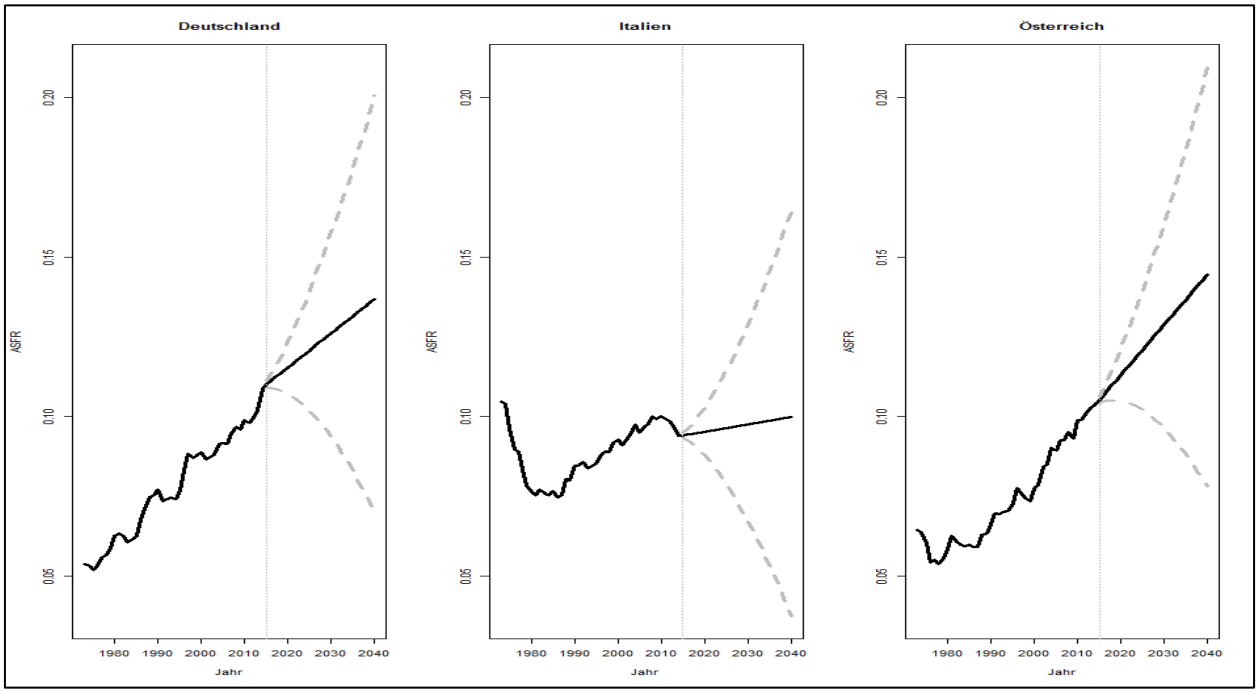
Auf Basis der Hauptkomponentenprognose lassen sich in der Folge durch die Rücktransformation von (30) aus den Prognosewerten der HK die Prognosewerte aller ASFR erhalten:

$$\boldsymbol{\Phi} = \boldsymbol{\Pi E^{-1}} \tag{31}$$

wobei $\boldsymbol{\Pi}$ die Matrix der Prognose der erwarteten Hauptkomponentenwerte von 2015 bis 2040 darstellt. $\boldsymbol{E^{-1}}$ ist die Inverse der Matrix der Eigenvektoren und $\boldsymbol{\Phi}$ schließlich die aus Multiplikation der beiden genannten Matrizen entstehende Matrix der Erwartungswerte für die Prognose der

ASFR bis 2040. Da die Prognose der Erwartungswerte allein nur wenig aussagekräftig ist, da ein Erwartungswert in einem solchen Zusammenhang mit einer asymptotischen Wahrscheinlichkeit von 1 nicht eintreffen wird, sollte zusätzlich die Unsicherheit der Prognose quantifiziert werden. Auch dies lässt sich über die HK bewerkstelligen. Da die HK untereinander unkorreliert sind, können durch simultane und unabhängige Computersimulationen der HK für jede HK separat Prognoseintervalle geschätzt werden. Für die HK wird dazu eine hinreichend große Anzahl an Pfaden simuliert. Der Autor hat für jede HK in diesem Fall 10 000 Pfade bis 2040 simuliert. Für HK 1 basiert die Simulation z.B. auf (27). Dabei wird wiederum auf die Stationaritätsannahme zurückgegriffen. Es wird dabei in jeder Periode der aus der Gleichung resultierende Erwartungswert gebildet und dazu eine Zufallsziehung aus einer normalverteilten Zufallsvariable mit der Standardabweichung gleich dem empirischen Wert in der jeweiligen Zeitreihe durchgeführt. Werden, analog zu (30), die Matrizen der Simulationsergebnisse der HK jeweils mit $\boldsymbol{E^{-1}}$ multipliziert, entstehen somit 10 000 simulierte Pfade für alle ASFR. Aus diesen können empirisch die Quantile für die gewünschten Grenzen der Prognoseintervalle gewonnen werden.

Abbildung 8: Prognostizierte ASFR für 30-jährige mit 80%-Prognoseintervallen



Quelle: Eigene Berechnung und Darstellung

Abbildung 8 illustriert beispielhaft die 80%-Prognoseintervalle für die drei Zeitreihen der ASFR der 30-jährigen in Deutschland, Italien und Österreich.

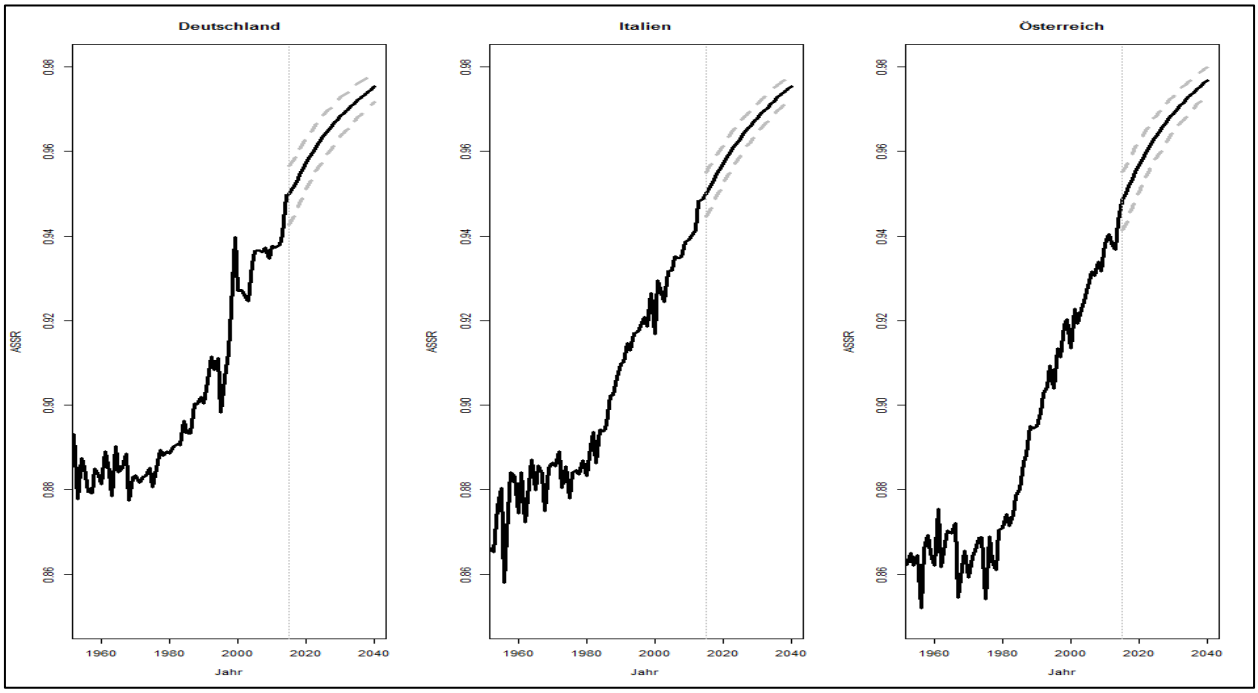
Nach einem sehr ähnlichen Schema lassen sich ebenfalls die Altersspezifischen Überlebensraten (ASSR) für die drei Länder berechnen, sodass auf die Berechnung an dieser Stelle nicht mehr so intensiv eingegangen werden soll. Während in der allgemeinen Betrachtung gerne auf Mortalitätsraten eingegangen wird, gilt die Analyse hier den ASSR, da diese für die Bevölkerungsfortschreibung von höherer Bedeutung sind. Da Überlebensraten, anders als Fertilitätsraten, grds. empirisch ausschließlich Werte über null und unter eins einnehmen, empfiehlt sich für die Prognose dieser anstelle einer direkten Nutzung der ASSR eine indirekte Prognose über den Logit der ASSR. Eine Logistische Transformation einer ASSR *s* lässt sich wie folgt durchführen:

$$logit(s) = \ln\left(\frac{s}{1-s}\right) \tag{32}$$

Die Transformation bewirkt, dass die neuen Variablen unbeschränkt sind, während die zugrunde liegenden ASSR nicht außerhalb des offenen Intervalls (0;1) prognostiziert werden können (Vgl. Johnson 1949: 157-158). Es muss nach der Simulation lediglich beachtet werden, dass die simulierten Werte noch durch den inversen Logit rücktransformiert werden müssen, um die Simulationswerte für die ASSR zu erhalten.

In dieser Studie wurden, ähnlich wie bei den ASFR, internationale Daten zur Sterblichkeit genutzt, um die gemeinsamen Trends in der Mortalität aufzufangen und die Datenbasis zu vergrößern. Im Detail wurden aus der Human Mortality Database, Eurostat und Datenlieferungen des Statistischen Bundesamtes altersspezifische Mortalitätsraten sowie Daten zu Bevölkerungszahlen und Sterbefällen nach Geschlecht aufgeteilt für 18 europäische Staaten extrahiert und aus diesen Daten die ASSR berechnet, bevor die Logit-Transformation für die ASSR durchgeführt wurde. Die Methode ergibt auf Basis von zwei Hauptkomponenten beispielhaft für 80-jährige Männer in Deutschland, Italien und Österreich die folgende Prognose mit 95%-Prognoseintervallen.

Abbildung 9: Prognostizierte ASSR für 80-jährige Männer mit 95%- Prognoseintervallen



Quelle: Eigene Berechnung und Darstellung

5 Fazit, Limitationen und Ausblick

Primäres Ziel dieses Beitrages war die Vorstellung der Hauptkomponentenanalyse und ihrer Anwendungsmöglichkeiten. Auf Basis der Hauptkomponenten ließen sich beliebige alters- und geschlechtsspezifische Größen ohne große Verzerrung und inklusive Quantifizierung der Stochastizität in Form von Prognoseintervallen kalkulieren. Das wurde an den Beispielen altersspezifischer Fertilitätsraten 30-jähriger Frauen in Deutschland, Italien und Österreich illustriert. Das zugrundeliegende Modell bezog sich auf altersspezifische Fertilitätsraten in den Altern 13-54 für Frauen aus allen 16 Ländern Süd- und Mitteleuropas, die eine Bevölkerungszahl von mindestens vier Millionen Einwohnern aufweisen. Daher kreiert das Modell simultan für all diese Länder und ASFR Prognosen. Als weiteres Anwendungsbeispiel wurden die Ergebnisse eines Modells zur Prognose altersspezifischer Überlebensraten für 18 westeuropäische Länder für 80-jährige Männer in den genannten drei Ländern illustriert. Generell ist auch an dieser Stelle zu erwähnen, dass der vorgestellte Modellansatz durchaus, wie gezeigt, international erweitert werden kann, als auch auf regionalere Ebenen heruntergebrochen werden kann. Es sei jedoch betont, dass die Modelle le-

diglich historische Trends aufnehmen, sodass radikale Veränderungen in Form von Strukturbrüchen nicht abgedeckt werden, insofern die Historie nicht ähnliche Vorkommnisse aufweist. Weiterhin sollte beachtet werden, dass die vorgestellten Ansätze durchaus methodisch korrekte Prognoseintervalle generieren, diese können jedoch, speziell bei längeren Prognosezeiträumen, sehr breit werden, was in erster Linie der relativ geringen Datenmenge geschuldet ist, die für demografische Prognosen grundsätzlich nur verfügbar ist. Der Modellierer kommt dementsprechend nicht umher, eine qualitative Beurteilung der erhaltenen Ergebnisse vorzunehmen und im Falle der Fertilität exogene Variablen einzubeziehen. Generell eignet sich der Modellrahmen sehr gut an, um parametrische Funktionen für langfristige Prognosen anzupassen. Die vorgestellten Modelle in dieser Arbeit waren daher nicht abschließend zu verstehen, sondern gaben nur einen Überblick über die Anwendungsmöglichkeiten der Methodik.

Als Einschränkung ist zu betonen, dass die Methodik für Migrationsprognosen nicht einfach zu übernehmen ist. Das liegt zum einen daran, dass die Datenverfügbarkeit in ihrer Korrektheit deutlich von der Datenqualität in den Bereichen Fertilität und Mortalität abfällt. Zudem ist die Stochastizität bei Migrationsflüssen erheblich größer (in Bezug auf Deutschland besonders bei der Immigration), da diese häufig in Schüben kommen und stark abhängig von wirtschaftlichen und gesellschaftlichen Trends sind. Das verdeutlicht sich z.B. in den hohen Massen an Menschen, die in den letzten Jahren aus Syrien und dem Irak fliehen (ähnliche Entwicklungen waren Anfang der 90er Jahre bei den Unruhen in Jugoslawien und dem Zerfall der Sowjetunion zu beobachten) und der daraus entstandenen lawinenartigen Migrationen aus anderen armen Ländern nach Europa, besonders nach Deutschland und Nordeuropa. Daher müssen bei Migrationsprognosen externe Faktoren einbezogen werden, was auf einer erweiterten Datenbasis sowie einer Erweiterung der Modelle durch qualitative Modellparameter passieren kann.

Literatur

Bell, William; Monsell, Brian 1991: Using Principal Components in Time Series Modeling and Forecasting of Age-Specific Mortality Rates. Vortrag auf der Jahrestagung der American Statistical Association. 18.-22. August 1991: Atlanta. In: 1991 Proceedings of the Social Statistics Section: 154-159.

Carter, L.R.; Lee, R.D. 1992: Forecasting demographic components: Modeling and forecasting US sex differentials in mortality. In: International Journal of Forecasting 8(3): 393-411.

Chatfield, Christopher; Collins, Alexander 1980: Introduction to Multivariate Analysis. Chapman & Hall.

Deschermeier, Philipp 2015: Die Entwicklung der Bevölkerung Deutschlands bis 2030 – ein Methodenvergleich. In: Vierteljahresschrift zur empirischen Wirtschaftsforschung des Instituts der deutschen Wirtschaft Köln 42(2): 97-111.

Dickey, D.A.; Fuller, W.A. 1979: Distribution of the Estimators for Autoregressive Time Series With a Unit Root. In: Journal of the American Statistical Association 74 (366): 427-431.

Engle, Robert 1982: Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation. In: Econometrika 50(4): 987-1007.

Eurostat. European Union. Abgerufen unter http://ec.europa.eu/eurostat/data/database [14.04.2016].

Fuller, W.A. 1996: Introduction to Statistical Time Series. 2nd Edition. Wiley Series in Probability and Statistics.

Glaister, Stephen 1984: Mathematical Methods for Economists. 3rd Edition. Basil Blackwell.

Graves, Spencer 2015: Package ´FinTS´. URL: ftp://cran.r-project.org/pub/R/web/packages/FinTS/FinTS.pdf.

Greene, W.H. 2012: Econometric Analysis. 7th Edition. Pearson.

Härdle, Wolfgang; Myšičkova, Alena 2005: Stochastic Forecast for Germany and its Consequence for the German Pension System. Humboldt Universität zu Berlin. SFB 649 “Economic Risk”. Diskussionspapier 2009-009.

Handl, Andreas 2010: Multivariate Analysemethoden: Theorie und Praxis multivariater Verfahren unter besonderer Berücksichtigung von S-PLUS. Springer.

Hannan, E.J.; Quinn, B.G. 1979: The Determination of the Order of an Autoregression. In: Journal of the Royal Statistical Society. Series B (Methodological) 41(2): 190-195.

Human Fertility Collection. Max-Planck-Institut für demografische Forschung (Deutschland) und Vienna Institute of Demography (Österreich). Abgerufen unter www.fertilitydata.org [14.04.2016].

Human Fertility Database. Max-Planck-Institut für demografische Forschung (Deutschland) und Vienna Institute of Demography (Österreich). Abgerufen unter www.humanfertility.org [14.04.2016].

Human Mortality Database. University of California, Berkeley (USA) und Max-Planck-Institut für demografische Forschung (Deutschland). Abgerufen unter www.mortality.org [31.08.2016].

Hyndman, Rob 2017: Package ´forecast´. URL: https://cran.r-project.org/web/packages/forecast/forecast.pdf.

Instituto Nacional de Estatística. Abgerufen unter https://www.ine.pt/xportal/xmain?xpid=INE&xpgid= ine_base_dados [14.04.2016].

Johnson, N.L. 1949: Systems of Frequency Curves generated by Methods of Translation. In: Biometrika 36(1/2): 149-176.

Keilman, Nico; Pham, D.Q. 2000: Predictive Intervals for Age-Specific Fertility. In: European Journal of Population 16(1): 41-66.

Kwiatkowski, Denis; Phillips, P.C.B.; Schmidt, Peter; Shin, Yongcheol 1992: Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. In: Journal of Econometrics 54: 159-178.

Lee, R.D. 1993: Modeling and forecasting the time series of US fertility: Age distribution, range, and ultimate level.

Simon, C.P.; Blume, Lawrence 1994: Mathematics for Economists. 1st Edition. W.W. Norton & Company, Inc.

Shumway, R.H.; Stoffer, D.S. 2011: Time Series Analysis and Its Applications: With R Examples. Springer.

STAT-TAB. Statistik Schweiz. Abgerufen unter https://www.pxweb.bfs.admin.ch [14.04.2016].

Statistisches Bundesamt 2005: Bevölkerung und Erwerbstätigkeit. Wiesbaden.

Statistisches Bundesamt 2015: Bevölkerung Deutschlands bis 2060: 13. Koordinierte Bevölkerungsvorausberechnung. Wiesbaden.

Statistisches Bundesamt 2016: Gestorbene: Deutschland, Jahre, Geschlecht, Altersjahre. Abgerufen unter https://www-genesis.destatis.de [24.08.2016].

Statistisches Bundesamt 2016a: Bevölkerung ab 31.12.1952 nach Alters- und Geburtsjahren: Bund. Daten auf Anfrage zur Verfügung gestellt [17.03.2016].

Trapletti, Adrian; Hornik, Kurt 2017: Package ´tseries´. URL: https://cran.r-project.org/web/packages/tseries/tseries.pdf.

Vanella, Patrizio; Pascariu, Marius 2017: A Principal Component Model for Forecasting Age- and Sex-Specific Survival Probabilities in Western Europe until the Year 2070. Vortrag auf der Jahrestagung des Deutschen Vereins für Versicherungswissenschaft. 15.-16. März 2017: Berlin.

Weidner, Wiltrud; Vanella, Patrizio; Zuchandke, Andy 2015: Die Entwicklung der Kfz-Zulassungen in Deutschland: Eine Prognose und Implikationen für die Kraftfahrtversicherung. In: Zeitschrift für die gesamte Versicherungswissenschaft 104(4): 365-387.