# 4

# Seasonal Analysis of Death Counts in the United States

## 4.1 Introduction

Demographers — as probably most other empirical researchers — prefer working with rates rather than with pure counts: growth rates, birth rates, death rates, transition rates, etc. The advantage is obvious: While count models rely only on the actual event of interest, rate models take also the units into account which are exposed to this event (e.g. person-years lived). Unfortunately, exposures are often not available. For example in the case of historical demography, the number of deaths by age and sex is regularly available. What is frequently absent, however, is the number of people who were alive (and therefore exposed to the risk of dying) in that particular age and sex. Also for the analysis of seasonal mortality, we are often faced with the situation to have death counts available but no exposures.

One way to avoid this problem is to estimate the exposures. Donaldson and Keatinge [77], for example, obtained the daily population in their study of winter excess mortality in southeast England "by linear interextrapolation from the 1981 and 1991 censuses". Also Kunst et al. [209] used linear interpolation for population estimates in their time-series analysis on the influence of outdoor air temperature on mortality in The Netherlands. Another solution in the case of absent exposures is to use only events. For those count models, it is not necessary to estimate any exposures. Typically, those studies assume an underlying Poisson process in the data like the analyses of seasonal variation in mortality in Scotland and in The Netherlands [121, 235].

The latter approach is clearly less desirable if exact exposures are available. If this is not the case, it is open to discussion whether an estimated population at risk is more favorable than pure counts. Especially in the case of seasonality studies, there are many problems associated with estimating seasonal populations (=exposures), as pointed out by Happel and Hogan [140].[1]

[1] It should be noted that Neale [271] already mentioned the problem of estimating monthly population counts in 1923.

This chapter presents an analysis of seasonality based on pure death counts in the United States from 1959 to 1998. Vaupel [381] once remarked that demographers should use the best possible data to study a certain phenomenon. Working with death counts as the best possible data seems to be contradictory at first sight as Scandinavian population registers, for example, offer exact event counts and precise exposure times. The *quality of the data* is, however, only one side of the coin of *best data*: it is equally important to take care of the *content of the data*. Small, egalitarian countries such as Denmark and Sweden with one common climate are less desirable than the US when one's aim is to study the impact of social factors on seasonal mortality. Thus, the "Multiple Cause of Death"-Public-Use-Files we used for the United States provide such a data-source: every individual death since 1959 is publicly available, broken down by various characteristics. The wealth of having almost 80 Mio. individual records available makes it possible to study selected causes of death for the whole period since the late 1950s across a wide age-range. More details of the data are explained in Section 4.3.

Besides the sheer amount of information, the lack of research on seasonal mortality in the United States during the last 25 years has been another reason to choose this country. Studies on seasonal mortality focused on European countries during the last 25 years. For the US, this topic has not been investigated since the late 1970s [231, 316, 319, 324, 325]. The only exception being regional studies (e.g. 199, 285) and one study on deaths from coronary heart disease by Seretakis et al. [340]. Solely, Feinstein [102] examined overall mortality in the United States recently. One important finding was that the "seasonalities of deaths have been increasing over the years [...] for older people and decreasing for younger people" [102, p. 485].

This was quite surprising. With the improved chances of people attaining high ages since the 1970s [378], we would have expected that elderly people were also better able to withstand environmental stress (i.e. cold in winter) with improvements in general living conditions.

## 4.2 Research Questions

There is ambivalent evidence for differences in seasonality of mortality for women and men. Some studies surprisingly found no differences for seasonality for this main determinant of mortality while others discovered remarkable differences between women and men in their seasonal mortality patterns with men showing larger seasonal fluctuations than women [98, 121, 262, 302, 419]. Therefore we decided to conduct all subsequent analyses for women and men separately.

- **Period & Cause of Death.** Do we find support for Feinstein's result of increasing seasonality for the elderly over time? Is it possible to detect different patterns for all cause mortality and selected groups of causes of death?

- **Age & Cause.** Previous studies have shown an increase in seasonality with age for various countries [251, 268, 302]. Can these findings be replicated in the US for all cause mortality and for selected causes of death?
- **Region & Period.** It is argued in the literature that socio-economic progress in general and the widespread use of central heating and air conditioning decreased seasonal fluctuations in deaths [188, 251]. We expect decreasing seasonality over time. However, regions with a high differential between winter and summer temperatures should have benefited more than regions with a moderate climate.
- **Region & Age.** How important is the region where you are living for the development of seasonal mortality? Is an assumed increase with age in seasonality of deaths larger in regions where one faces higher environmental stress than in other regions?
- **Education, Age & Cause of Death.** The question how socio-economic status — a major general mortality determinant [374] — affects seasonality in deaths is still unanswered. Few studies argue that lower social groups are disadvantaged [e.g. 79, 147]; most others found no social gradient [214, 215, 342]. Our analysis focuses on the question whether people with higher education face lower seasonality in deaths.
- **Marital Status & Age.** Another major factor in mortality research is marital status, usually showing that married people have lowest (overall) mortality. Typically, married people have lower mortality risks throughout their life courses than single, widowed or divorced persons. Men's differences are larger than women's [129, 163, 223]. These differential mortality risks are usually explained either by a protection effect or by a selection effect [125, 223]. In the case of seasonal mortality, a protection effect can be imagined in several directions: people who are married can pool their financial resources and have therefore not only better access to medical care, but can also afford a higher quality of housing which is a major determinant in avoiding cold-related mortality as previous studies have shown [e.g. 245]. While this causal pathway could be also captured by education as a proxy for socio-economic status, marital status may also work in another direction: in comparison to single, widowed and divorced people, married women and men are most likely not living alone. In the case of an emergency, the spouse is usually present to organize help. Nevertheless, no research has been published so far on the potential impact of marital status on seasonal mortality.

## 4.3 Data

Our analysis uses the "Multiple Cause of Death"-Public-Use-Files for the years 1959–1998 published by the "US Centers for Disease Control and Prevention" (CDC). We downloaded the data from 1968–1998 from the "Inter-university Consortium for Political and Social Research" (ICPSR) at

http://www.icpsr.umich.edu/. Data for previous years have been kindly provided by the "Program on Population, Policy and Aging" at the Terry Sanford Institute for Public Policy at Duke University, NC.

We included only deaths at ages 50 and higher, because we wanted to focus on adult mortality. At younger adult ages, the number of deaths in certain age-groups for selected causes of deaths are too few to obtain robust estimates. The data consist of more than 77 Mio. individual death records. Each of the records contains information on the sex of the individual, month and year of death, age at death. For our analysis, we also extracted information on the cause of death, state of residence and state of occurence, and several social variables. Figure 4.1 gives an overview on the availability of these variables in our data over time. The following subsections explain how we divided and coded the data for our analysis.

### 4.3.1 Cause of Death

Table 4.1 outlines which ICD codes we used to extract the information for our selected causes of death. ICD is the abbreviation for "International Statistical Classification of Diseases and Related Health Problems" from the World Health Organization (WHO). This coding scheme gives mandatory instructions how the cause of death has to be coded. During its existence, the ICD underwent several revisions. While ICD-10 is the current revision, ICD-7, ICD-8, and ICD-9 were in use in the United States during our observation period. ICD-7 was used until 1967; between 1968 and 1978 ICD-8 was the valid coding scheme; from 1979 until 1998 deaths in the United States were coded according to ICD-9.

Table 4.2 gives an overview about the actual number of deaths for each cause. In addition, we have given information about the contribution of each cause to all deaths for the whole time-series, for the first five years, and the last five years to highlight vaguely any time trends. In the column "Winter/Summer Ratio" we divided winter deaths (January–March) by summer deaths (July–September) to find out whether our selected causes show a considerable seasonal difference in mortality. We did not give an extra-column for a test for seasonality. All causes of death presented here have passed Hewitt's nonparametric test for seasonality with significant values ($\rho = 0.0130$) [150, 395]. This indicates that all causes examined show a pattern where the six highest values of a year and the six lowest values of a year are not mixed but appear in separate halves of the year. Most people died of cardiovascular diseases during our observation period, with almost 32 Mio. deaths. In conjunction with neoplasms, cerebrovascular and respiratory diseases almost 80% of deaths are covered. Despite the regularities in the ordering of the months (i.e. significant results for Hewitt's test), the extent of seasonality differs remarkably: On average (=All Causes), the number of summer deaths is exceeded by winter deaths by roughly 16%. Neoplasms, not surprisingly, show relatively small fluctuations (1.6%), whereas respiratory diseases have 62%

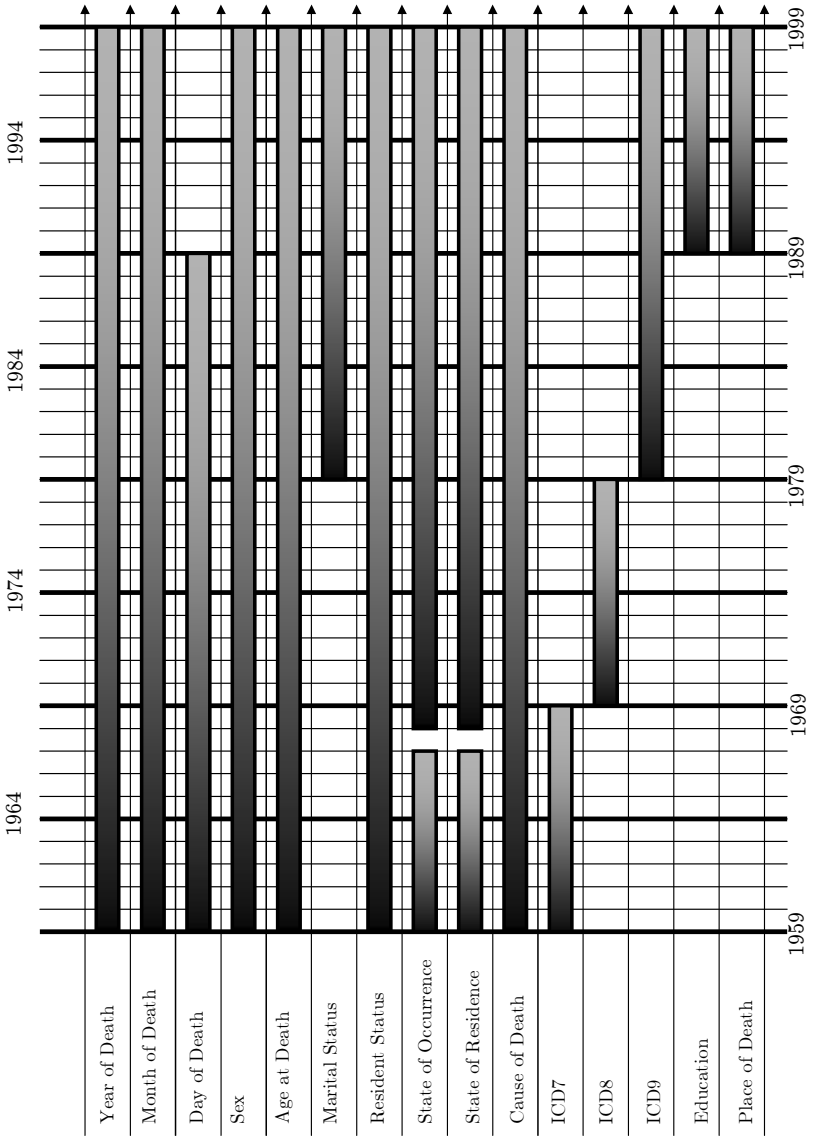



**Fig. 4.1.** Availability of Variables in Our Data-Set over Time

more deaths in winter than in summer. The leader in that respect is influenza with a value of 27.762 (i.e. almost an excess of 2,700%).

Although the most remarkable changes in the cause of death structure over time are usually associated with the "epidemiological transition" [281] and the vanishing of tuberculosis [402] in the 20$^{\text{th}}$century, the proportions of the leading causes of death have not remained constant during recent decades either.

**Table 4.1.** Coding Scheme for Selected Causes of Death

| **Cause of Death** | **ICD-Codes** | | |
|---|---|---|---|
| | **ICD-7** | **ICD-8** | **ICD-9** |
| **Cardiovascular Diseases** | 400–468 | 390–429 | 390–429 |
| | | | 440–459 |
| | | 440–458 | 557 |
| *IHD* | — | 410–414 | 410–414 |
| **Neoplasms** | 140–239 | 140–239 | 140–239 |
| **Cerebrovascular Diseases** | 330–334 | 430–438 | 430–438 |
| **Respiratory Diseases** | 240; 241 | 460–519 | 460–519 |
| | 470–527 | | |
| *Asthma* | 241 | 493 | 493 |
| *Influenza* | 480–483 | 470-474 | 487 |
| *Pneumonia* | 490–493 | 480–485 | 480–483 |
| | | | 486 |
| *Bronchitits* | 500–502 | 490–491 | 490–491 |
| **Diabetes Mellitus** | 260 | 250 | 250 |
| **Infect. & Parasit. Dis.** | 001–138 | 001–136 | 001–139 |
| *Tuberculosis* | 001-019 | 010–019 | 010–019 |
| **Liver Cirrhosis** | 581 | 571 | 571 |

With the exception of IHD (1968–98), all causes of death are covered for the period 1959–1998.

Figure 4.2 gives an overview of how seven major causes have changed during our observation period. For "both sexes", "women", and "men" there are two columns each, showing the cause-of-death spectrum for the first (1959–63) and last (1994–1998) five years, respectively, covered in our dataset. Cardiovascular diseases remain the leading cause of death (see also Table 4.2) — although the contribution shrunk for both sexes from 45% to 35%. Similarly, also cerebrovascular diseases lost in relevance between the late 1950s and the late 1990s. Almost 12% of all people died from that group of diseases between 1959–63, whereas in the years 1994–98 only 7% died of it. Net "winners" in this respect are mainly malignant neoplasms (17% → 22%) and respiratory diseases (7% → 9%). Diabetes Mellitus and "Infectious and Parasitic Diseases" also gained in relevance, however their overall share is comparatively small (Diabetes Mellitus: 1.84% → 2.66%; Infectious and Parasitic Diseases: 1.24% → 2.80%). It is interesting to note that influenza and hypothermia — two causes of death which are often associated with winter excess mortality — make up only a negligible part of all deaths (influenza: 0.04%; hypothermia: 0.02%). These small proportions, however, might mask the real impact of these diseases. For example, it is well-known that "[i]nfluenza epidemics cause deaths additional to those registered as being due to influenza, such as deaths caused by arterial thrombosis"[78, p. 90].

**Table 4.2.** Number of Deaths, Proportion and Seasonal Pattern of Selected causes of Death

| **Cause of Death** | **# of deaths 1959–1998** | **Proportion of Cause, 1959–1998** | **Proportion of Cause, 1959–1963** | **Proportion of Cause, 1994–98** | **Winter/ Summer Ratio** |
|---|---|---|---|---|---|
| **All Causes** | 77,640,423 | 100.00% | 100.00% | 100.00% | 1.157 |
| **Cardiovascular Diseases** | 31,926,214 | 41.12% | 44.97% | 34.84% | 1.206 |
| *IHD (1968-98)* | 17,422,235 | 27.96% | 29.23% | 21.10% | 1.211 |
| **Neoplasms** | 16,335,426 | 21.04% | 16.84% | 22.15% | 1.016 |
| **Cerebrovascular Diseases** | 7,055,237 | 9.09% | 11.84% | 6.88% | 1.197 |
| **Respiratory Diseases** | 5,688,290 | 7.33% | 5.76% | 9.70% | 1.624 |
| *Asthma* | 153,338 | 0.20% | 0.31% | 0.24% | 1.295 |
| *Influenza* | 107,048 | 0.14% | 0.28% | 0.04% | 27.762 |
| *Pneumonia* | 2,356,304 | 3.03% | 3.45% | 3.50% | 1.789 |
| *Bronchitits* | 174,048 | 0.22% | 0.27% | 0.14% | 1.663 |
| **Diabetes Mellitus** | 1,599,351 | 2.06% | 1.84% | 2.66% | 1.187 |
| **Infectious and Parasitic Diseases** | 1,226,575 | 1.58% | 1.24% | 2.80% | 1.151 |
| *Tuberculosis* | 160,856 | 0.21% | 0.62% | 0.06% | 1.201 |
| **Liver Cirrhosis** | 1,080,511 | 1.39% | 1.27% | 1.10% | 1.098 |

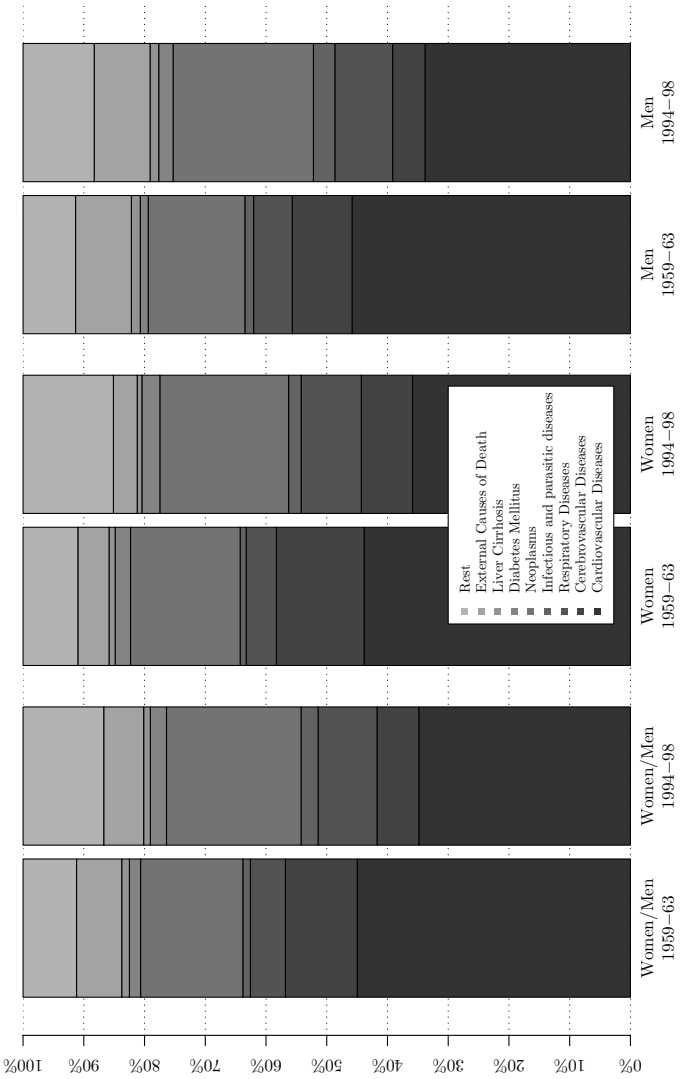



**Fig. 4.2.** Changes in the Cause of Death Composition of Adult Deaths in the United States Between 1959–63 and 1994–98 by Sex

### 4.3.2 Education

The variable education has been included since 1989. The original data are given as a two-digit code indicating years of education. We followed the recoding advice in the coding manual with one exception: we included two

additional categories which indicate whether a person has finished elementary school (8 years of education), dropped out of elementary school (less than 8 years) or has received no formal education at all (0 years). All other categories remained the same and have been given meaningful labels. The categories, their labels and the corresponding numbers of death broken down by sex are given in Table 4.3.

**Table 4.3.** Number of Deaths Broken Down by Sex and Level of Education

| Code | Meaning | Deaths | | | |
|---|---|---|---|---|---|
| | | Women | | Men | |
| | | Counts | % | Counts | % |
| 0 | No formal education | 105,462 | 1.0 | 108,348 | 1.0 |
| 1 | Elementary School Dropout | 846,138 | 7.9 | 975,483 | 8.7 |
| 2 | Finished Elementary School | 1,390,687 | 12.9 | 1,229,727 | 10.9 |
| 3 | High School Dropout | 1,197,240 | 11.1 | 1,229,727 | 10.9 |
| 4 | Finished High School | 3,746,633 | 34.8 | 3,594,343 | 31.9 |
| 5 | College Attendance | 1,120,953 | 10.4 | 1,147,500 | 10.2 |
| 6 | College Degree or more | 857,895 | 8.0 | 1,264,765 | 11.2 |
| 7 | Not Stated | 1,502,673 | 14.0 | 1,549,632 | 13.8 |
| | $\Sigma$ | 10,767,681 | 100.0 | 11,249,981 | 100.0 |

Finishing high school was the most common level of education achieved by both sexes (women: 34.8%; men: 31.9%). Although our decomposition in 7 categories is relatively detailed, enough people remain even in the smallest group "no formal education" with more than 100,000 deaths for each sex.

### 4.3.3 Marital Status

Data on marital status are available since 1979. To make comparable analyses on the impact of social factors by age, we restricted our analysis to the years 1989–98, the same period as for education. In the official codebooks six categories are given which have been converted to five: never married / single, married, widowed, and divorced remained the same. The category "not stated on certificate" has been merged together with "not stated". This residual category comprises less than one percent of each sex (♀: 0.3%, ♂: 0.7%). In contrast with the variable "education", the cell frequencies differ remarkably between women and men. Most notable are the differences for married and widowed women and men. This is the result of the higher life expectancy of women. It is more likely for women at the end of their lives to be widowed than for men.

**Table 4.4.** Number of Deaths Broken Down by Sex and Marital Status

| Code | Marital Status | Deaths | | | |
|---|---|---|---|---|---|
| | | Women | | Men | |
| | | Counts | % | Counts | % |
| 1 | Never Married, Single | 935,504 | 8.7 | 1,536,393 | 13.7 |
| 2 | Married | 2,820,570 | 26.2 | 6,487,584 | 57.7 |
| 3 | Widowed | 6,102,184 | 56.7 | 2,011,515 | 17.9 |
| 4 | Divorced | 879,450 | 8.2 | 1,136,594 | 10.1 |
| 9 | Not Stated | 29,973 | 0.3 | 77,895 | 0.7 |
| | $\Sigma$ | 10,767,681 | 100.0 | 11,249,981 | 100.0 |

### 4.3.4 Region

Various studies have shown that countries with relatively harsh climatic conditions and cold winters (e.g. Canada, Sweden) show less winter excess mortality than countries with warm or moderate climate such as Portugal, Spain or the UK [135, 147, 252]. It is argued that people in colder regions are better able to protect themselves against adverse environmental conditions. One disadvantage of previous studies was that these results were based on cross country analyses. The data from the United States provide an excellent framework to analyze seasonal mortality in different climatic regions within one country. For our regional analysis we followed the state groupings given in the original coding manuals which resemble different climatic regions. Our slightly adapted division of states is presented in Table 4.5. In its original version the states Alaska and Hawaii belonged to the group "Pacific". In our analysis, these two states have been examined separately. Figure 4.3 makes it easier to locate the coding of the regions geographically. This classification resembles in most cases the "Köppen Climate Classification". In some cases, however, the regional classification does not describe states with similar meteorological conditions. For example, Arizona and Montana in the "Mountain-Group" differ considerably in their climate. Special care should therefore be taken for the interpretation if estimations from the "Mountain" and from the "Midwest" show exotic results.

We refer to the actual "state of occurrence", i.e. the state/region where the death has happened. "State of residence" is given in the data as well. In our analyses by region we only included those deaths where state of residence and state of occurrence were in the same regional division excluding the impact of "snowbirds" [140].[2] The loss of data is relatively minor. More than 98% of all deaths happened in the same region as the place of residence of the deceased.

[2] People who are seasonally migrating — usually to warmer regions during the cold season — are sometimes labeled "snowbirds" in the literature.

**Table 4.5.** Coding of Regions by State

| Code | Region | States | | |
|---|---|---|---|---|
| 1 | New England | Connecticut | Maine | Massachusetts |
| | | New Hampshire | Rhode Island | Vermont |
| 2 | Middle Atlantic | New Jersey | New York | Pennsylvania |
| 3 | Midwest | Illinois | Indiana | Iowa |
| | | Kansas | Michigan | Minnesota |
| | | Missouri | Nebraska | North Dakota |
| | | Ohio | South Dakota | Wisconsin |
| 4 | South Atlantic | Delaware | D.C. | Florida |
| | | Georgia | Maryland | North Carolina |
| | | South Carolina | Virginia | West Virginia |
| 5 | South Central | Alabama | Arkansas | Kentucky |
| | | Louisiana | Mississippi | Oklahoma |
| | | Tennessee | Texas | |
| 6 | Mountain | Arizona | Colorado | Idaho |
| | | Montana | Nevada | New Mexico |
| | | Utah | Wyoming | |
| 7 | Pacific | California | Oregon | Washington |
| 8 | Alaska | Alaska | | |
| 9 | Hawaii | Hawaii | | |

### 4.3.5 Known Data Problems

Generally speaking, the "US Multiple Cause of Death"-Public-Use-Files provide a very good basis for research. Nevertheless, there are some real and some potential pitfalls in the data which will be briefly outlined here as well as the approaches used to tackle them.

ICD Revisions: During our observation period, three revisions of the ICD were in practice in the US (ICD-7, ICD-8, ICD-9). If one is not careful, the introduction of a new revision is prone to result in sudden shifts in the number of deaths. An illustrative example is Asthma. While ICD-7 was used, this disease (ICD-7 code: 241) belonged to the group of "Allergic, endocrine system, metabolic and nutritional diseases" (ICD-7 Codes: 240–289). Since the eighth revision, Asthma (ICD-8 code: 493) is one of the "diseases of the respiratory system" (ICD-8: 460–519). Therefore particular care was taken in reconstructing the time-series. Besides consulting the original coding schemes, the following procedures have been undertaken to obtain time-series with a maximum of quality:

- The first step was to plot the data to discover any breaks or otherwise strangely behaving characteristics in the data. As pointed out by Cleveland: "Data display is critical to data analysis. Graphs allow us to explore data to see the overall pattern and to see detailed behaviour;

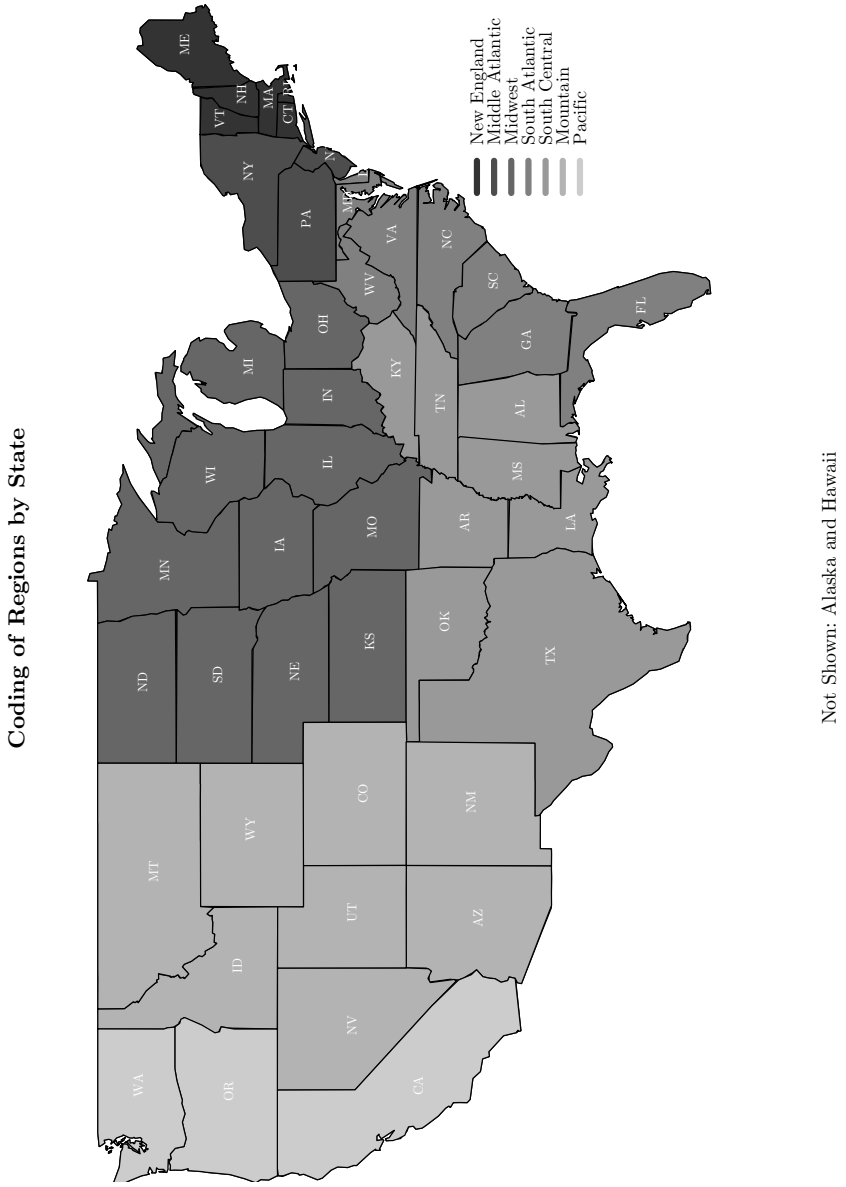



**Fig. 4.3.** Coding of Regions by State

no other approach can compete in revealing the structure of the data so thoroughly" [49, p. 5].
- Articles and monographs by Jacques Vallin and France Meslé were consulted (e.g. [259, 375]) who are probably *the* experts on reconstructing time series of causes of death.
- Several articles on seasonal mortality give details about the ICD codes they used for a particular cause [e.g. 98, 209]. This was valuable in finding "hidden" causes such as asthma mentioned before. The scope of some articles covered more than one ICD coding scheme. Marshall et al. [246], for example, give the ICD codes for Coronary Heart Disease for ICD-8 and ICD-9. Articles like this facilitated the transition from one ICD revision to the next.
- The statistical software package Stata with its search facilities for ICD codings (`icd9 search` and `icd9 lookup`) allowed to find all possibilities for a certain disease which would otherwise remain undetected.
- Vladimir Shkolnikov, Michael Bubenheim, Sigrid Gellers-Barkmann, Rembrandt Scholz and Markéta Pechholdová from the "Laboratory for Demographic Data" at the Max Planck Institute for Demographic Research in Rostock, Germany, have given valuable advice and suggestions for the reconstruction of the time-series.

The Year 1972: In the year 1972, the Multiple Cause of Death Public Use File contained only a 50% sample of all deaths. We simply multiplied all deaths by a factor of 2 to circumvent this problem.

The Years 1987 & 1988: We discovered a sudden drop in death counts by plotting annual deaths for selected causes for the years 1987 and 1988. After checking several possibilities as a cause, we found out that only the first 44 US states (in alphabetical order) had been included for those two years. Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, and Wyoming were missing. We tackled this problem by estimating the contribution of those states for the year 1986 and 1989 for our respective analysis (e.g. for sex, age group and educational level). With those two values we made a linear interpolation of what we would expect for the years 1987 & 1988. We then multiplied the actual counts for those states with a factor to obtain the expected number of deaths. Of course, this does not solve the problem perfectly. Nevertheless, we believe that this approach yields more satisfactory results than, for example, leaving out these 7 seven states for all analyses.

## 4.4 Methods

### 4.4.1 Model Requirements

The data used in this project have specific features that we need to take into account when selecting the appropriate models for analysis: the employed

methods should allow for the count character of the data, without requiring information on the corresponding exposures. Covering a period of four decades of remarkable changes in mortality, especially at older ages, the data show considerable variation in the overall trend, both between different causes of death but also between different age-groups within the same cause. Thus, appropriate models have to allow for a flexible specification of these different trend features. We do not know how the trend and the seasonal component changes with age and/or over time. Therefore we do not want to impose any specific parametric model upon our data but rather use data-driven, non-parametric techniques to estimate our components. Last but not least, we would like to allow for overdispersion in our models as this "is the norm in practice and nominal dispersion the exception" [249, p. 124–125]. As shown in Chapter 3 (Measuring Seasonality), previously existing methods such as X-11, STL, ... were unable to extract the exact trend and the exact seasonal component. Therefore, a new method has been developed which is presented in the following sections to fulfill these requirements.

### 4.4.2 The Model

#### Basic Model Specification

Let $t$ denote the underlying time variable which can represent calendar-time or age. For matters of convenience in this explanation, $t$ represents calendar-time. The corresponding number of deaths, corrected for the different lengths of months, is denoted $y_t$. Our model resembles several characteristics from the well known field of *generalized linear models* (GLMs):

Distribution: We assumed that the $y_t$ follow a Poisson distribution with parameter $\mu_t$. Thus $\mathrm{E}(y_t) = \mathrm{Var}(y_t) = \mu_t$. The Poisson distribution is usually regarded as "the benchmark model for count data"[41, p. 3].

Link Function: Similar to the setting of GLMs, we relate $\mu_t$, which are the expected values of $y_t$ to a stimulus matrix via a link function. In our case, the stimulus matrix is time (or age) and transformations of it. While other link functions are also possible for Poisson distributed data (for example, the square-root- or the identity-link, see [389]), we use the canonical/default choice of a log-link.

The model we are estimating is:[3]

$$\ln \mu_t = f(t) + \sum_{l=1}^{L} \left\{ f_{1l}(t) \sin\left(\frac{2\pi\, l}{12}\, t\right) + f_{2l}(t) \cos\left(\frac{2\pi\, l}{12}\, t\right) \right\}. \qquad (4.1)$$

[3] It should be pointed out that the development of this model is based on an idea of Dr. Jutta Gampe. The model was implemented in strong collaboration between her and the author.

The model is estimated in a similar manner as a GLM. The main deviation are the parameters which are estimated. In the GLM setting, *one parameter* is estimated for each column in the covariate matrix. In our model, these scalars are replaced by functions. These functions are indicated by $f(t)$, $f_{1l}(t)$ and $f_{2l}(t)$ in Equation 4.1. The component $f(t)$ describes the varying trend in the level of counts — due to changing exposures and overall changes in mortality. The seasonal fluctuations are modeled with the latter two terms in the equation. In the most simple case with $L = 1$, two seasonal functions $f_{11}(t)$ and $f_{21}(t)$ are estimated, resulting in one(!) smoothly changing annual fluctuation. If $L = 2$, a semi-annual swing is added. Theoretically, it is possible to add higher frequencies. It is doubtful it will make sense, though, if $L \geq 3$. These kinds of models have been termed *varying coefficient models* by Hastie and Tibshirani [145]. "In contrast to the GLM, where the regression coefficients [...] are assumed to be constant, [...] this model accommodates situations in which one or more of the coefficients are allowed to vary smoothly (interact) over [...] time or space"[89, p. 760].

### Technical Digression: Nonparametric Estimation of Smooth Trends Varying Coefficients

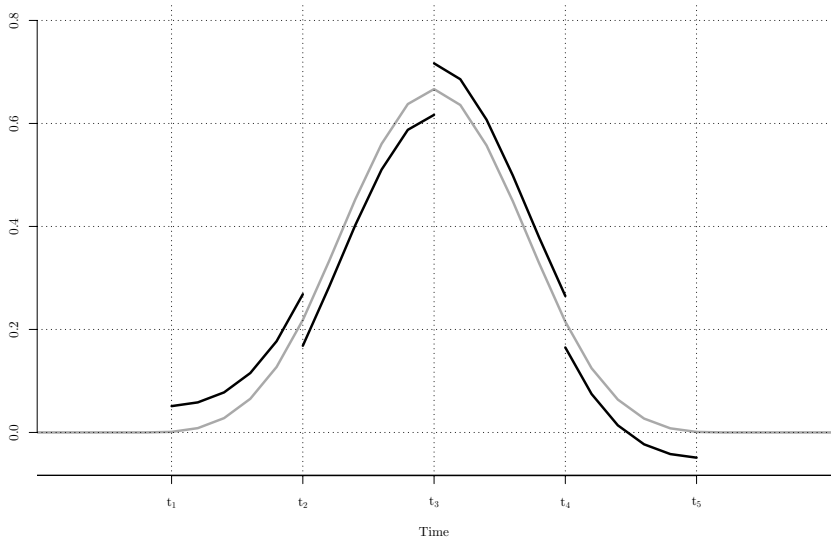
The following section, until page 101, represents a technical digression.[4] The aim is to show how a function like $f(t)$, $f_{1l}(t)$ or $f_{2l}$ is actually estimated. The equations in this section (Equations 4.2 and 4.3) are not directly linked to Equation 4.1.

We assumed that $f(t)$, $f_{1l}(t)$ and $f_{2l}$ are smoothly changing over time (or age). In a recent paper, Eilers and Marx [89] showed that such models, which they termed *GLASS* (Generalized Linear Additive Smooth Structures), can be estimated via $P$-Spline smoothing. This technique belongs to the family of nonparametric smoothers. $P$-Splines are cubic $B$-Splines being used as regression bases with a roughness penalty on their regression coefficients. $B$-Splines are made of polynomial pieces connected with knots. Please see Figure 4.4 for a graphical explanation.[5] In our case of cubic (degree $q = 3$) $B$-Splines, each $B$-Spline consists of $q+1 = 4$ polynomial pieces, as indicated by the four segments in gray. Each of these polynomial pieces is of degree $q = 3$. These polynomial pieces are connected at $q = 3$ inner knots $(t_2, t_3, t_4)$. At those knots, the spline function as well as the $q - 1 = 2$ derivatives of the neighboring polynomial pieces are continuous. The $B$-Splines are positive on a domain of $q + 2 = 5$ knots. This corresponds in Figure 4.4 to the range from $t_1$ to $t_5$ on the time-axis; everywhere else they are zero [87, 132]. These $B$-Splines are bell-shaped and resemble a Gaussian density (=density of a Normal distribution) [89] without the smoothing problems when regression bases are derived

[4] As this approach is novel, it is appropriate to include it in the main text instead of putting it into the appendix.

[5] An extensive discussion of $B$-Splines (definition, basic properties, ... is given in [67].

from a normal distribution. For example, Gaussian smoothers cannot fit a straight line as they are not locally defined but from $[-\infty; \infty]$ resulting in a "Gaussian ripple" [86]). Such an example is given in Apppendix C on page 183.




Adapted from Eilers and Marx [86].

**Fig. 4.4.** Construction of One $B$-Spline

With these cubic $B$-Splines as regression bases, we are working in the well known area of linear regression. The smoothed function is found by minimizing $S$ in Formula (4.2) via the traditional OLS-fitting. In this equation, $y$ represents the response vector, $B$ the matrix of covariates (=our $B$-Splines) and $\alpha$ their respective regression coefficients.

$$S = |y - B\alpha|^2 \tag{4.2}$$

Figure 4.5 shows cubic $B$-Splines "in action" to smooth artificial data.[6] In the lower part of each of the four panels, you see cubic $B$-Splines which are close to normal densities as postulated. From left to right and from top to bottom, the number of $B$-Splines is increasing. The upper part of each

[6] It might be interesting to note that the use of cubic $B$-Splines is relatively widespread: For example, the software to design the letters of this text (METAFONT) used some cubic $B$-Splines to have smooth and visually appealing shapes [200].

panel shows scatterplot of the data and a line. This line is the result of the smoothing using the cubic $B$-Splines as regression bases.

One can easily see:

- The higher the number of $B$-Splines, the closer (and "wigglier") the smoothed curve is to the data.
- The lower the number of $B$-Splines, the smoother is the curve.

The problem one faces now is to find an optimally smoothed curve. If the curve is too smooth, important characteristics of the data are not caught. If the curve is too wiggly, the data are overfitted, i.e. we include more complexity into the model than what is actually desirable. There is no golden standard for choosing the optimal number of $B$-Splines and therefore of regression parameters.[7] One could follow a subjective approach to determine the optimal number of parameters. Although it may sound repulsive to the "objective" scientist, "[i]t may well be that such a subjective approach is in reality the most useful one" [132, p. 29]. We are following another approach outlined by Eilers and Marx [87] as no all-purpose scheme existed to choose the optimal number of splines automatically. The idea is simple: Building on works of O'Sullivan [283] and Reinsch [306], they proposed to choose a relatively large number of cubic $B$-Splines which would normally result in over-fitting. To prevent this fallacy, a penalty is put on the regression coefficients. More specifically, a penalizing constant is multiplied with the second derivative of the regression coefficients.[8] The previous optimization problem (in Formula 4.2) changes to Formula 4.3:

$$S^* = |y - B\alpha|^2 + \lambda |D_2\alpha|^2 \quad \text{, where } D_2\alpha = \Delta^2\alpha \tag{4.3}$$

The iterative procedure to optimize $S^*$ has been described in [89]. Figure 4.6 shows the impact of how a change in the penalizing parameter $\lambda$ affects the smoothness of the curve. In all of the nine panels we see the same artificial data as in Figure 4.5. The number of cubic $B$-Splines has been set to a relatively high level, which would result in over-fitting if the regression coefficients were not penalized. With a $\lambda$-value of 0.01 in the upper-left panel, the weight of the penalty-term is relatively negligible, resulting in the expected overfitted, wiggly curve. The higher the $\lambda$-values (from left to right and up / down), the smoother the curve gets. While the upper two graphs are definitely too close to the data, the last curves are — without any doubt — too smooth

[7] As we are actually using regression parameters, the term "non-parametric models" might be misleading. Eilers and Marx [87] pointed out that "anonymous models" is preferrable as parameters are estimated. They simply have no scientific interpretation.

[8] Eilers and Marx note that the second derivative has been used since "the seminal work on smoothing splines by Reinsch (1967)", however, "[t]here is nothing special about the second derivative; in fact, lower or higher orders might be used as well" [87, p. 91].

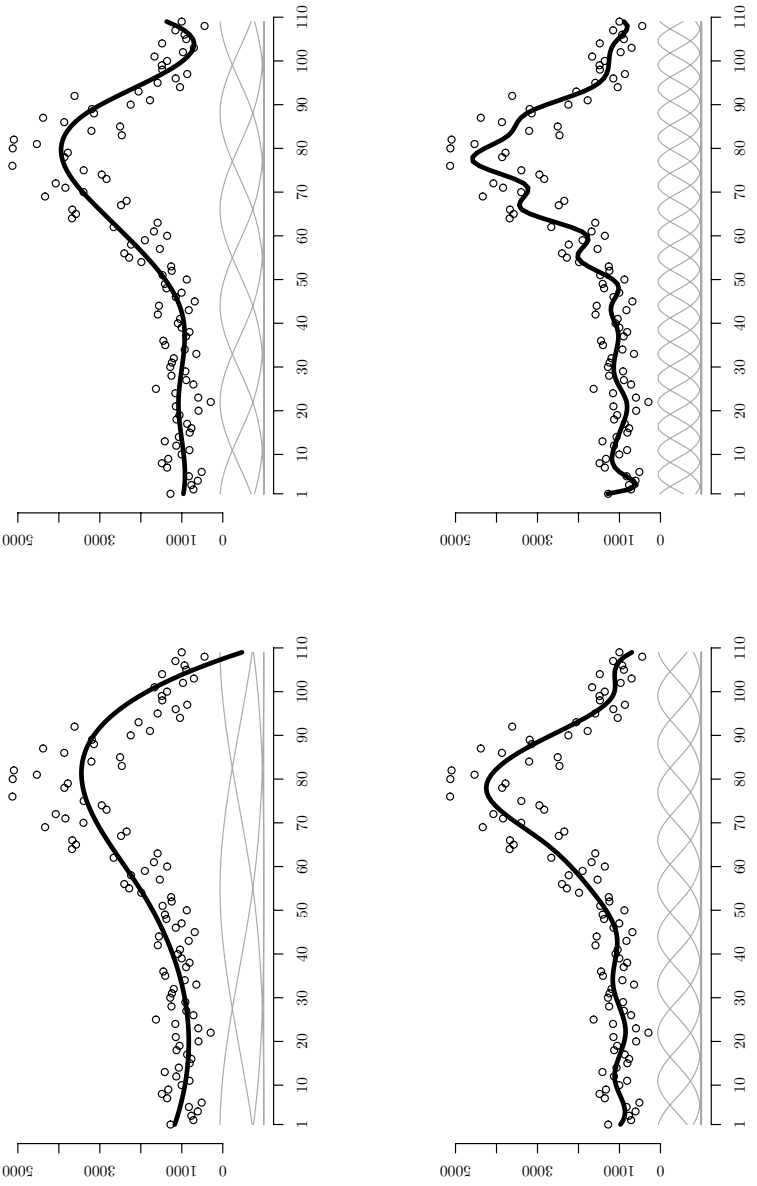



**Fig. 4.5.** Smoothing of Artificial Data Using Different Numbers of Cubic $B$-Splines as Regression Basis

with the outcome that important characteristics of the data are not captured. Ultimately, the smoothed curve tends to become a horizontal line for $\lambda \to \infty$.

There are several strategies to find the optimal value of $\lambda$, for example cross-validation. We followed the path of Eilers and Marx [87] and used the Akaike Information Criterion (AIC). Put in a nutshell, the AIC corrects the fit of the model for the number of parameters involved in the model's estimation.

*P*-Splines have several useful properties which makes Eilers and Marx [87, p. 98] "believe that *P*-splines come near to being the ideal smoother." For example, their foundation in linear regression and the generalized linear model makes them easy to understand and use. Also the lack of unwanted boundary effects favors *P*-Spline smoothers instead of other smoothing methods.[9] An exhaustive comparison of various smoothing methods, their properties and their respective pros and cons are found in [88].

### Overdispersion & Smoothing Parameter Selection

After initial experiments, we discovered that our data violated one of the key assumptions of the Poisson distribution which we were using; As stated in Formula 4.4, the mean and the variance are characterized by the same parameter (we denoted the parameter by $\mu$ as the standard choice; $\lambda$ is already in use for the smoothing penalty parameter).

$$\mathrm{E}(y_t) = \mathrm{Var}(y_t) = \mu_t \tag{4.4}$$

As mentioned before, this assumption of nominal dispersion is relatively strong. Regularly, one observes *overdispersion* in practical applications. Overdispersion is defined as $\mathrm{E}(y_t) = \mu_t < \mathrm{Var}(y_t)$. This case, where the variance exceeds the mean, can arise for various reasons [22]:

- if the rate $\mu_t$ is not constant within a chosen time unit $t$ (*time dependence in the rate*)
- if the number of events in a time-interval depends on the number of previous events (*contagion*).[10]
- in the case of *unobserved heterogeneity*, i.e. there are covariates not entered into the model which affect the number of counts.

All of them are likely for our analysis of death count data — especially *unobserved heterogeneity*. We can certainly expect two sources of unobserved

[9] In the case of smoothing with a polynomial, it can not be excluded that some values are estimated at the boundaries which do not make sense. For example, if a quadratic curve is used for smoothing, the fitted line points on both ends go either up or down although it is possible that the resulting values do not have any theoretical meaning (e.g. lifetimes smaller than zero).

[10] It should be noted that already Greenwood and Yule stated in 1929 [133, p. 276] "the problem of the distribution arising when the chance of a happening is affected by antecedent success or failure".

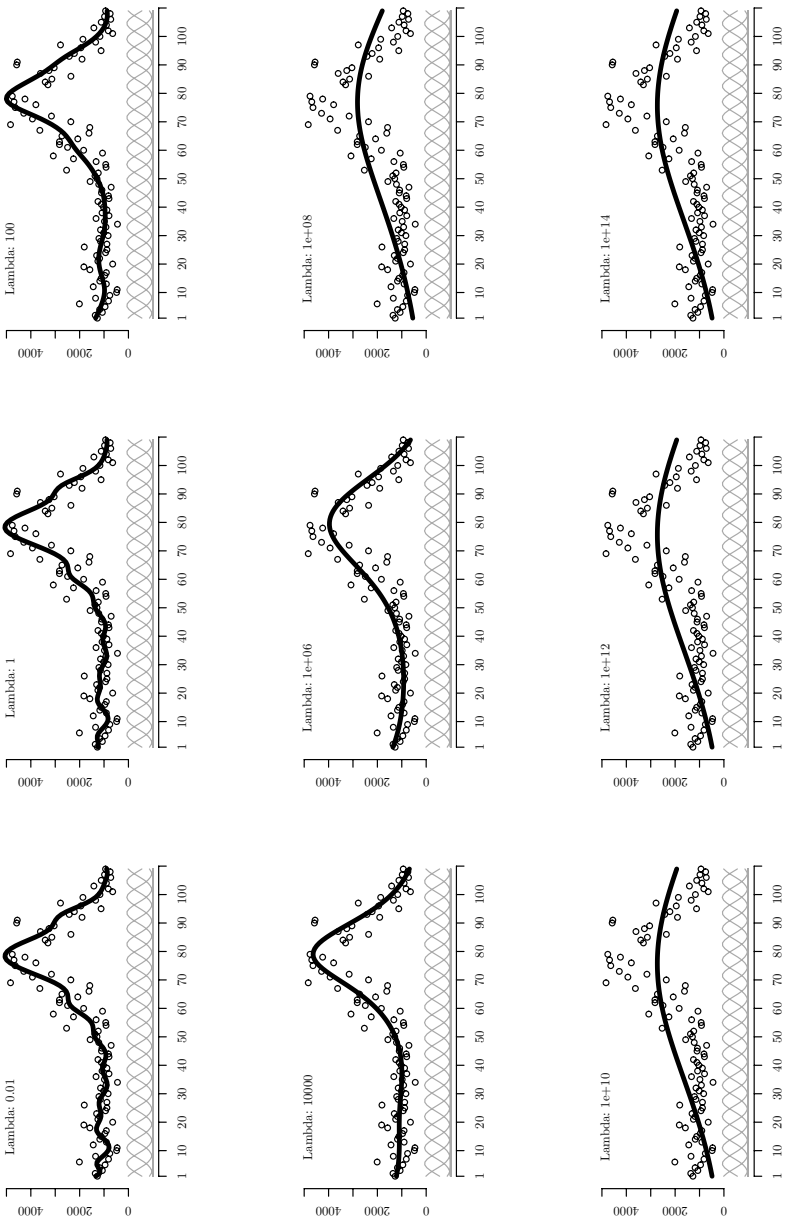



**Fig. 4.6.** The Impact of Changing $\lambda$-Parameters on the Smoothness of the Curve

heterogeneity in our data: one is due to the fact that the month of death is only a proxy-variable for the actual factors (e.g. temperature) which modulate the expected number of deaths $\mu_t$ seasonally. There is individual but unobserved heterogeneity in the risk of death for specific months across years. Secondly, even if we restrict the analysis to one sex, narrow age-groups, ..., people in these groups are heterogeneous with respect to other characteristics not included in the analysis.

Although we do not know what the actual reason of overdispersion is, there is a way to control for it. The typical approach still follows the suggestion of Greendwood and Yule [133] of assuming that the data follow a Poisson distribution, "but there is gamma-distributed unobserved individuals heterogeneity reflecting the fact that the true mean is not perfectly observed" [41, p. 71]. This modeling of a random effect for the mean with a gamma distribution leads to the Negative Binomial Distribution for the count [41, 160, 292].

The Negative Binomial Distribution is closely related to the Poisson Distribution as the following tabulation shows:

| **Distribution** | **Expected Value** | **Variance** |
|---|---|---|
| Poisson | $\mu_t$ | $\mu_t$ |
| Negative Binomial | $\mu_t$ | $\mu_t + \frac{\mu_t^2}{\theta}$ |

The estimator for the expected value remains the same: $\mu_t$. Using this parameterization of the variance as shown by Venables and Ripley [389], we can easily recognize that the Negative Binomial distribution depends simply on one more parameter called $\theta$. One could argue that the Negative Binomial Distribution is a generalization of the Poisson distribution by relaxing the term for the variance. We can model the Poisson case of nominal dispersion by letting $\theta \to \infty$. The other extreme of large overdispersion can be modelled by letting $\theta \to 0$.

The problem that arises is now: which $\theta$-value is to be chosen, as this parameter has to be entered into our model? The solution is found in the properties of the so-called *Pearson Residuals* in the Generalized Linear Model. They are defined as [see 249, p. 37]:

$$r_P = \frac{y - \mu}{\sqrt{\mathrm{V}(\mu)}} \quad \text{and in our case and notation:} \quad r_{P_t} = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{\mu}_t + \frac{\hat{\mu}_t}{\theta^\lambda}}},$$

where $y_t$ denotes the number of deaths at time $t$ (for the analysis by period), $\hat{\mu}_t$ represents the estimated value at time $t$ and Var is the estimated variance. This standardization of the *raw residuals* $(y_t - \hat{\mu}_t)$ results for an optimal model in large samples in $\mathrm{E}(r_{P_t}) = 0$ and $\mathrm{Var}(r_{P_t}) = 1$ [41, p. 141].

Our strategy for choosing the optimal model proceeds in the following steps.

1. We assume a grid of possible overdispersion parameters $\theta$.
2. For each given $\theta$:
   - we estimated all possible models with the given grid of all $\lambda$-permutations. In the simplest case when $L = 1$ in Equation 4.1, three separate $\lambda$s were estimated.[11]
   - we estimated the AIC from all models estimated in the previous step and chose the one with the minimum AIC value.
3. We iterated the previous step for all values of $\theta$.[12]
4. The outcome of the previous step was one "conditional optimal model" for each given $\theta$. Then, we calculated the Pearson Residuals for these "conditional optimal models". The one model where the variance of the Pearson Residuals was closest to 1 was then chosen to be the optimal model.[13]

Using simulated data, we compared our final model which incorporates overdispersion with a model which assumed data following a poisson distribution. This approach has the advantage that we know the various components that are entered into the model and can therefore check whether the two decomposition approaches return the same components we have entered into our simulated data. Figure 4.7 shows such a simulated example in a $3 \times 4$ panel. The left column displays the simulated data. The trend component (Figure 4.7 d) has been constructed by using a third-order polynomial. The seasonal component is linearly increasing (Fig. 4.7 g). We assumed a value of 10 for $\theta$ in the Negative Binomial Distribution which results in high overdispersion. This is reflected in the residuals as shown in Figure 4.7 j. Apart from the linear increase in the seasonal component, this model is equivalent to Model VII in Chapter 3 presented on page 78.

The middle column represents the optimal model, which has been estimated using our approach which incorporates unobserved heterogeneity. Figure 4.7 b shows the entered time-series which is equivalent to Figure 4.7 a. It can be clearly seen that the extraction of the trend (Fig. 4.7 e) and of the seasonal component (Fig. 4.7 h) mirrors the input data almost perfectly. As demanded from our model, the variance of the Pearson residuals should be 1 for the optimal model. Figure 4.7 k shows that our estimation is reasonably close enough with a value of 1.04. The right column exemplifies a mis-specified model. Although we used a Negative Binomial Model in the middle column

[11] If we had given 5 values for $\lambda_{Trend}$ which estimates the trend function $f(t)$, and also 5 values each for the penalty coefficient for the seasonal functions $f_{1l}(t)$ and $f_{1l}(t)$, we would have had to estimate $5 \times 5 \times 5 = 125$ models.

[12] If we had also given 5 possible values for $\theta$, $125 \times 5 = 625$ models would be required to be estimated.

[13] If the variance of the Pearson Residuals was not close enough to 1, we started again with step 1 with an increased grid. "Close enough to 1" for the variance of the Pearson Residuals was defined as: $0.99 < \mathrm{Var}(r_{P_t}) < 1.01$.

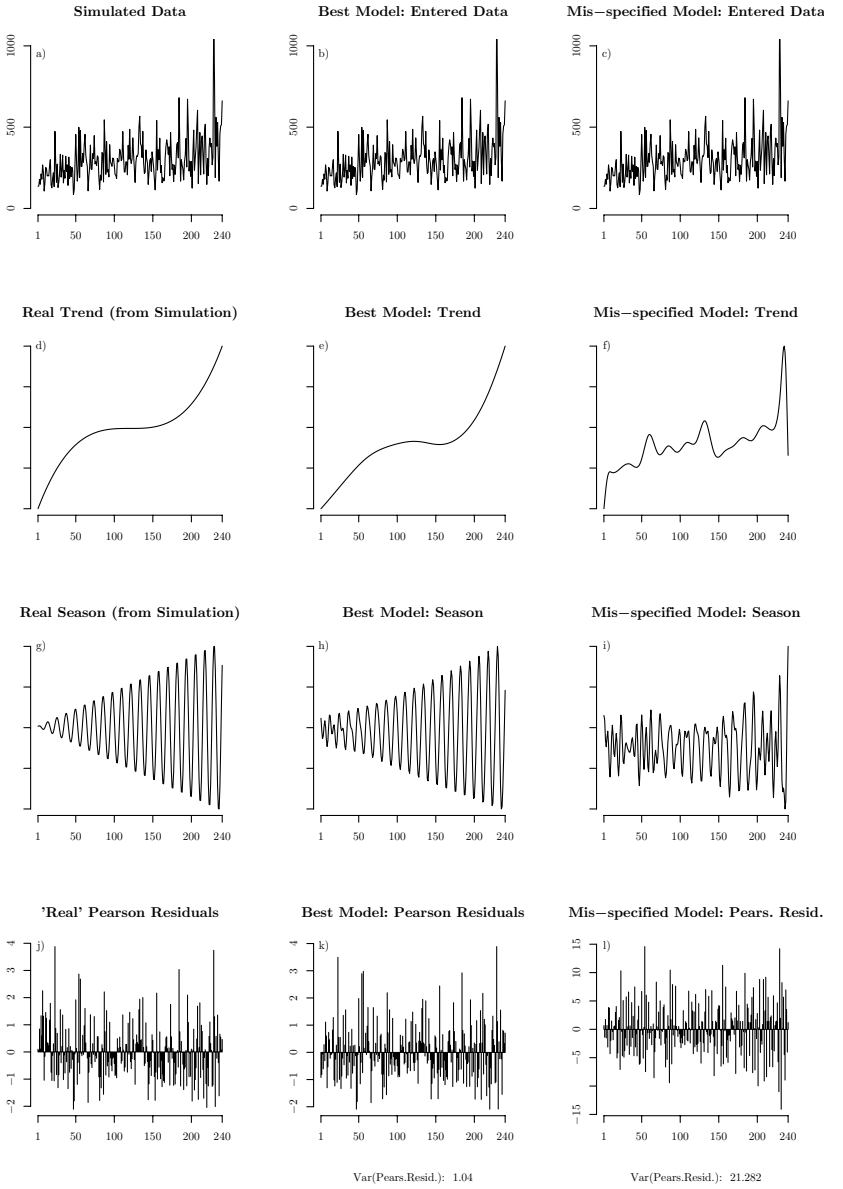



**Fig. 4.7.** Simulated Data, "Optimal" Model and a Mis-specified Model

as well, we have chosen a value for $\theta$ (= 9000), which approximates a Poisson Distribution. Without taking unobserved heterogeneity into account our model is helpless in estimating the trend (Fig. 4.7 f) and the seasonal component (Fig. 4.7 i). Not surprisingly, the variance of the Pearson Residuals in

the mis-specified model (Fig. 4.7 l) is far too high (21.282) for an expected value of 1.

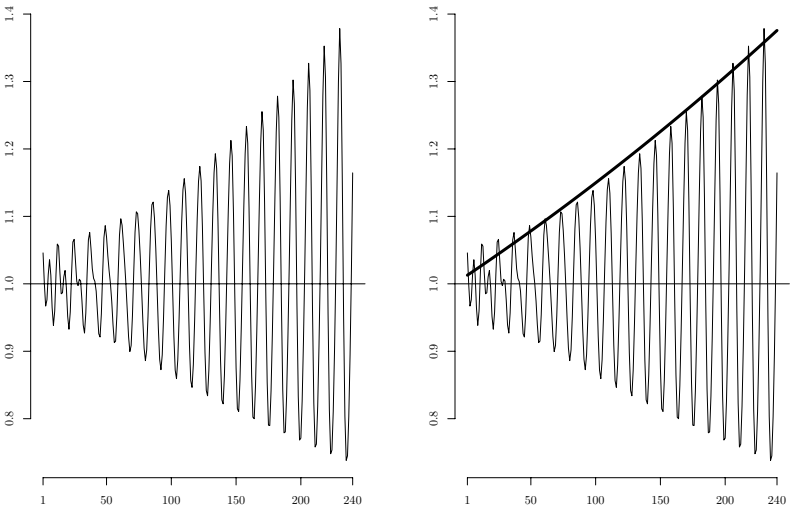



**Fig. 4.8.** Seasonal Component from Figure 4.7 and Its Amplitude

Based on this simulation study,[14] we concluded that our approach meets our requirements that an appropriate method should be able to work with overdispersed data having a flexible trend and a changing seasonal component.

The following section presents the results of our analysis for which we used the decomposition method outlined here. We are, however, only interested in a small part of the three components: the change in the amplitude of the seasonality over time (or age). This corresponds in mathematical notation to the smooth amplitude-modifying functions of the seasonal components in Equation 4.1 (page 96) [118]:

$$a_l(t) = \left(f_{1l}(t)^2 + f_{2l}(t)^2\right)^{\frac{1}{2}} \tag{4.5}$$

We use the resulting function $a_l(t)$ from Equation 4.5 and plot $e^{a_l(t)}$. Figure 4.8 explains this graphically. In the left panel our extracted seasonal component for an optimal parameter selection from Figure 4.7h is displayed. The difference between the left and the right panel is that in the latter we added the amplitude over time($e^{a_l(t)}$) of the seasonal fluctuations using a bold line.

[14] Of course, more simulation studies have been conducted. The one presented here should only serve as an example.

This line is used as indicator of the change in seasonality over time or age in the subsequent sections. A value of 1 corresponds, thus, to the case when no seasonality is present. These seasonality values should not be confused with the exponentiated regression coefficients known from event-history models and understood as relative risks. Values apart from 1 have no direct interpretation.

## 4.5 Results & Discussion

### 4.5.1 Seasonality by Period & Cause of Death

#### All Cause Mortality

Figure 4.9 shows the change in the amplitude for seasonality in deaths from all causes by 10-year-age-groups for the whole observation period from January 1959 until December 1998. The left panel illustrates results for women, whereas the right panel deals with men. For both sexes we see the same general trends: the older the people (=the darker the lines), the higher is the seasonal amplitude. Changes over age will be examined in subsequent parts of this section. Right now the focus is on changes over time. What we discover is some preliminary support for Feinstein's finding: Younger age-groups seem to have a constant or slightly decreasing trend as indicated by the dotted and dashed gray lines — especially for men. People who died at an age above 80 (dotted, dashed and solid lines), however, have to suffer from higher fluctuations in seasonality towards the end of the observation period.

With the progress made in survival chances — especially for older people — we would have suspected that people are better able to withstand environmental stress in recent times. A solution for this surprising finding is not straightforward. One has to keep in mind that "Seasonality for All Cause Mortality over Time, by Age-Group" is an aggregated outcome over several variables. Between 1959 and 1998 mean age at death, measured by $e_0$, rose from 73.24 years to 79.31 years for women (♂ 1959: 66.80 years; 1998: 73.53 years) in the United States [166]. Consequently, also the distribution of deaths within one 10-year-age-group shifted upwards. Among octogenarians, for example, the arithmetic mean for age at death increased from 83.68 years to 84.01 years for men (♀ 1959: 83.94 years; 1998: 84.54 years). This compositional effect might blur the "true" effect of changes in seasonality over time. We checked this problem by estimating seasonality for all cause mortality over time by single ages. The results (not shown here) resembled our findings for 10 year-age-groups: at least since the late 1970s, seasonalities are increasing for the elderly.

#### Selected Causes of Death

After ruling out the impact of the age composition, we decomposed the aggregated picture into selected causes of death and analyzed them separately for women and men by age-group as shown in Figure 4.10.

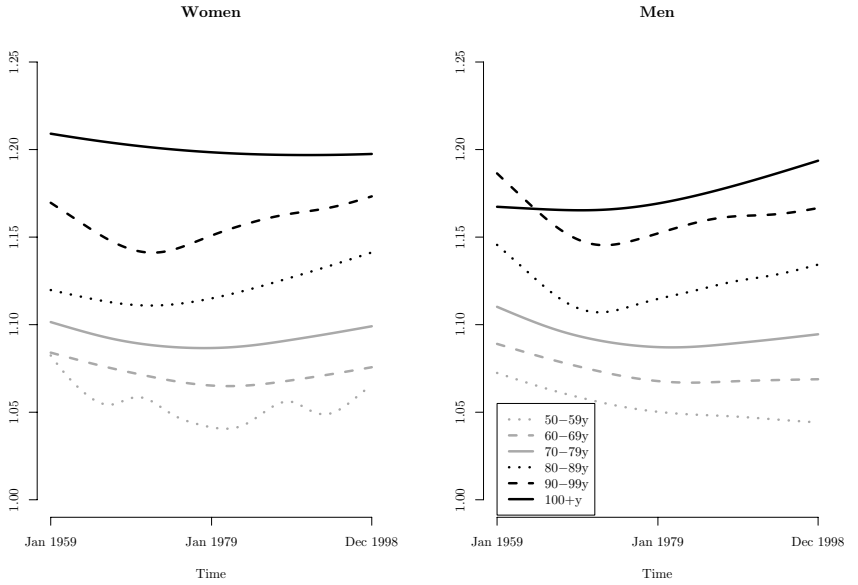



**Fig. 4.9.** Seasonality of All Cause Mortality over Time by Sex and Age-Group

Deaths from cardiovascular diseases are shown in the upper left panel for women and in the upper right panel for men. As this was and still is the leading group of cause of death, the fact that the two diagrams resemble the results for all cause mortality rather closely is not too surprising. Cerebrovascular diseases, as illustrated in the two panels in the middle row, are similar to the previous pictures for deaths from all causes as well as from cardiovascular diseases. The increasing trend for the elderly is even more obvious for this cause of death category. Apart from women who have died between 50 and 59 years of age from that cause (dotted gray lines), all seasonalities are increasing at least since the middle of the 1970s.[15]

If data problems can be excluded, there are always two strains of explanation which can be referred to when interpreting changes in populations [383]. First, there is a *real* difference in the variable of interest (seasonal susceptibility). In our context, this explanation would imply that people have become more susceptible to environmental stress over the years. One has to be careful with this interpretation, though: by looking at the changes in the amplitude we are using a relative measurement. An increase in the amplitude can either be

[15] The dashed gray line on the left, denoting deaths of women aged 60-69 years, represents an outlier. So far, it has been impossible to track down the source for this problem since several checks have been already conducted such as plotting the time series, looking for sudden changes in the number of deaths, etc. Also a completely new approach for estimation resulted in the same outcome.

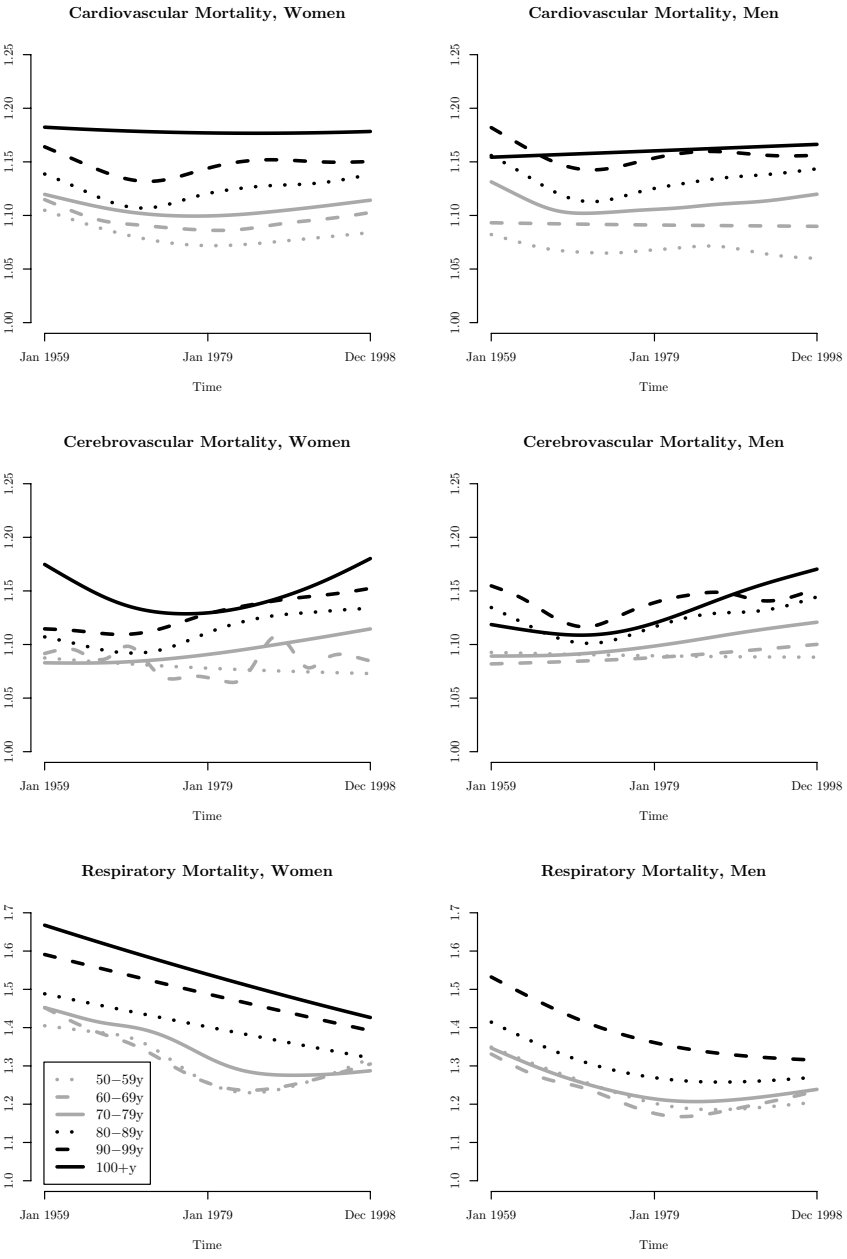



**Fig. 4.10.** Seasonality of Selected Causes of Death over Time by Sex and Age-Group

caused by a real increase in winter mortality or by a decrease in winter mortality over time with decreases in summer mortality at an even faster pace. The results would be the same: a larger seasonal amplitude in deaths/mortality by the end of the observation period than in the beginning. This conjecture finds support in the article of Seretakis et al. [340], who also found a decrease in deaths from coronary heart disease until the 1970s followed by a slight increase: "If the reversal is real, then it could reflect the increase in use of air-conditioning, which would have blunted the effects of occasional heat waves on coronary mortality" [340, p. 1014].
Secondly, however, there is the possibility that the change is influenced by a *compositional* difference. In the context of seasonal mortality fluctuations, it is possible that current progress against old-age mortality has the side-effect that nowadays even frail people can become relatively old. While in the past, frail individuals died early and left a relatively robust cohort of survivors who were coping well with environmental hazards in winter, frail people today may become older and are more susceptible in winter. This explanation could be, however, only applied to people at relatively advanced ages.

Not all causes of death show the same pattern over time. The two panels on the bottom of Figure 4.10 contain the development of seasonality over time for respiratory diseases. For both sexes we observe a decline in seasonality. While the decrease is almost linear for women at advanced ages, men's and "younger" women's seasonality shrank until the late 1970's, and has stalled since. Please note the different scale on the y-axis in comparison to cerebro- and cardiovascular diseases: seasonality for respiratory diseases was much higher in the past and still is. Although this gap has become smaller, the amplitudes in seasonal death fluctuations from respiratory diseases remain higher in comparable age-groups.
In univariate analysis of time-series it is always difficult to determine causal influences of external variables. It is, however, quite likely that improvements in housing conditions played a major part. While in the US in 1960 only half of all households were heated by gas or electricity, this proportion reached 82 percent in 1990 [372]. With these improved chances to heat the house properly, chances are decreasing for people to "catch a cold". The different pattern for women and men cannot be explained by this, though.

### 4.5.2 Seasonality by Age & Cause of Death

Of prime interest for demographers are not only death patterns for women and men over time but — maybe even more important — with age. Previous articles state that seasonality is increasing with age. However, the data used in many studies appear to be problematic [cf. 302, p. 199]: "In several studies, no distinction by age was made at all [13, 21, 319, 367]. If the factor age was taken into account, the highest included age or the beginning of the last, open-ended, age category was chosen at an age after which most deaths in a population occur [37, 62, 76, 97, 98, 164, 169, 188, 253, 369]. The oldest people

in the "Eurowinter"-study, for example, were 74 years old. Bull and Morton [37] merely made a binary distinction: younger than 60 years; 60 years and older. Thus, results from these studies may simplify or blur the relationship between age and seasonal fluctuations in mortality " [302, p. 199]. So far, there are only a few studies that have investigated seasonality in mortality or deaths into very high ages [251, 268, 302]. The highest ages that have been analyzed were centenarians and supercentenarians (110 years and older) in the study of Robine and Vaupel [309]. Regardless whether they calculated seasonality indices and ratios or log-odds, the typical outcome were higher seasonal fluctuations by the end of the lifespan than at middle ages. Even supercentenarians show higher excess winter mortality than centenarians, which indicates that also at those ages, the resistance against environmental hazards is decreasing [309].

### All Cause Mortality

Figure 4.11 gives a first impression how seasonality in deaths changes with age. The left panel shows seasonality for deaths from all causes for women where each solid line indicates a 10-year-calendar period. The right panel shows results from the respective analysis for men.

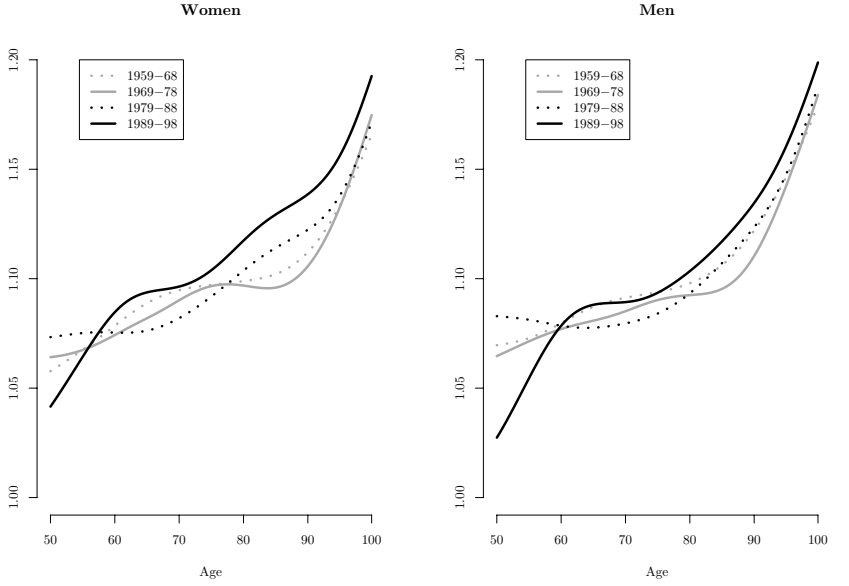



**Fig. 4.11.** Seasonality of All Cause Mortality by Sex and 10-Year-Calendar-Pperiod

The general trend for both sexes shows — as expected — higher seasonality with age. The increase is far from linear. We could make a distinction for women as well as for men by grouping the first three decades together (dotted gray: 1959–68; solid gray: 1969–78; dotted black: 1979-88) and contrast them with the last 10 years (1989–98 in solid black): Until age 80 the increase is relatively moderate. Then, at the highest ages, seasonality bends sharply upwards. The black solid line in both panels represents changes with age for the most recent decade in the analysis (1989–98). One can differentiate three stages: Compared to previous decades, seasonality is relatively low at age 50 and increases until age 60 where it is roughly on level terms. Between 60 and 75/80 years seasonality remains relatively constant. After age 80, seasonality in deaths from all causes is increasing, and shows higher values than in the past for the same ages.

**Selected Causes of Death**

To gain further insights, we decomposed the pattern for all causes again into the three major seasonal diseases. The results are shown in Figure 4.12 for cardiovascular (upper left & upper right panel), cerebrovascular (middle left & middle right panel), and respiratory diseases (lower left & right panel).

As we have seen previously for the analysis by calendar-time, seasonality of cardiovascular diseases matches seasonality from all causes almost perfectly. Especially for men during the most recent decade analyzed (1989–98), we recognize again the development of seasonality in three stages. While the age-range 60–65 marks also here the bending point from an increase in seasonality to a constant pattern, the age when the slope becomes steeper again is shifted to the right. Seasonality for cardiovascular deaths shows a strong upward tendency after ages 90–95. This three-stage-process is also repeated for cerebrovascular diseases with only slightly changing ages as turning-points. I would like to stress that the puzzling pattern is not the outcome of our model. If we had chosen a polynomial for our estimation procedure those unwanted boundary effects could have been implicit in the model as mentioned briefly in the end of Section 4.4.2. Using the *P*-Spline approach, though, has the advantage that "[b]oundary effects do not occur if the domain of the data is properly specified" [87, p. 98]. Excluding, thus, data problems, we propose an interaction between "real" changes in susceptibility and compositional changes due to mortality selection. Following the mortality model proposed by Robine [310], increasing mortality reflects vanishing resistance towards environmental hazards. The same should hold for seasonality: with increasing age, seasonal fluctuations should become larger as the human body becomes more and more susceptible to the detrimental effects of winter. At the same time, we observe a selection effect in mortality: "All populations are heterogeneous. [...] Some individuals are frailer than others, innately or because of acquired weaknesses. The frail tend to suffer high mortality, leaving a select subset of survivors. [...] As a result of compositional change, death rates increase more slowly with age

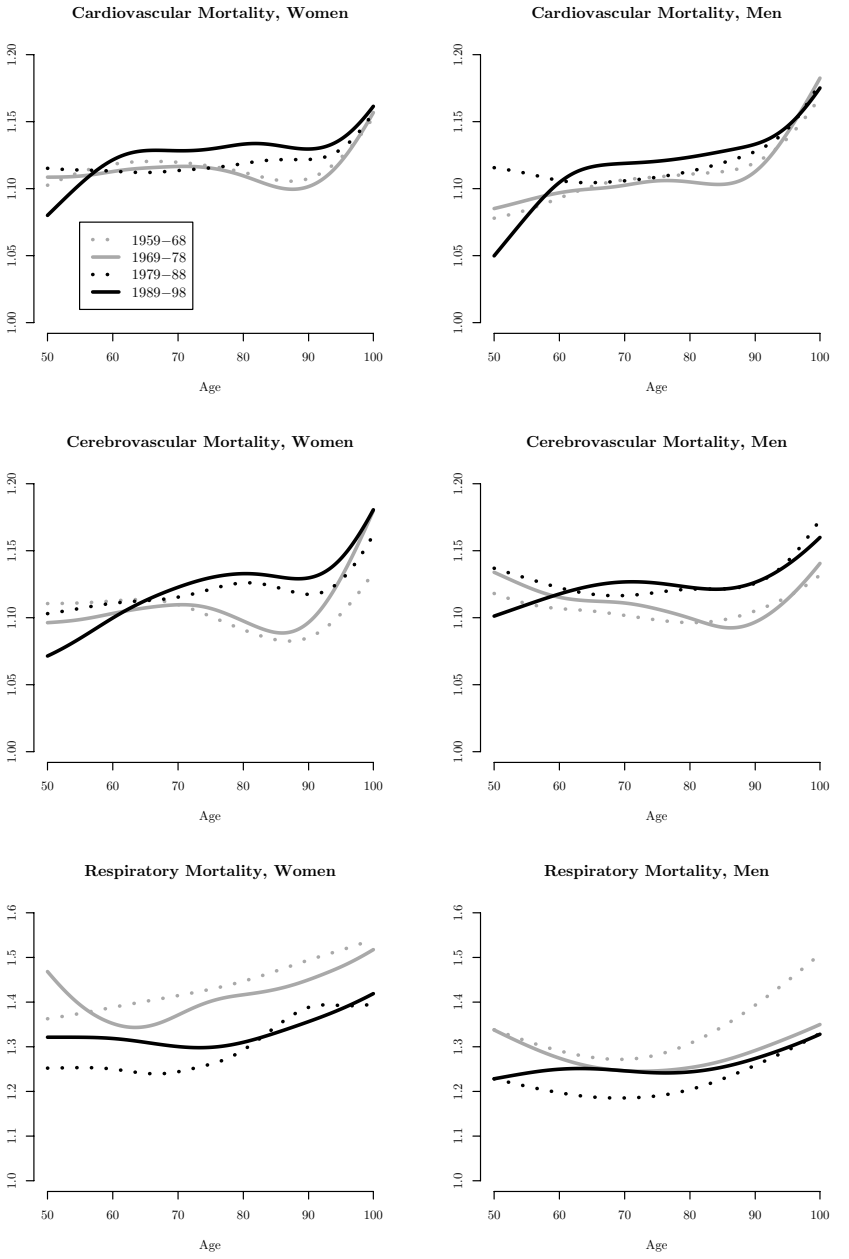



**Fig. 4.12.** Seasonality of Selected Causes of Death by Sex and 10-Year-Calendar Period

than they would in a homogeneous population." [384, p. 858]. This might also have a decreasing effect on the magnitude of seasonality in deaths. In our case we can argue that this selection effect is relatively weak before age 65, as not many people have died out of the population. At subsequent ages the push-factor for the seasonal amplitude (higher susceptibility) is balanced out by the pull-factor (mortality selection). This effect can be easily simulated following the concepts of Vaupel and Yashin [386]. Figure 4.13 shows one of the "ruses" selection effects can play: Our population consists simply of two sub-populations. The frail sub-population is getting seasonally more susceptible in a linear fashion (dotted, gray line). The more robust sub-population — as shown by the dashed, gray line — is relatively immune to stressful environmental conditions during winter into their late 80's. During the last few years of their lives, seasonality increases at a faster pace. We do not know who belongs to the robust group and who to the frail group. What we observe is the population level illustrated by the solid, black line.[16]

By this simple simulation with two subpopulations we can easily see that our observed outcome in Figures 4.11 and 4.12(upper 4 panels) could be generated by such a process. Further support can be drawn from these graphs by looking at the development over calendar time: the depressing impact of the selection effect is getting smaller over time. This could reflect the fact that in the past there were only relatively robust survivors in those higher age-groups, whereas nowadays people are reaching those ages who would not have been able to do so only 20 years earlier.

The lower two panels in Figure 4.12 show the change in seasonality with age for deaths from respiratory deaths. For this cause of death, we have not discovered a pattern as for the two previous causes. After a slight decrease for women as well as for men until age 65, seasonality increases steadily with age. The two panels also give support for the previous finding in Figure 4.10 (page 109): Over the course of the observation period, seasonal fluctuations have become smaller in more recent decades as indicated by the four plotted lines. Thus, improvements in general living conditions seemed to help in reducing the annual cold-related death toll due to infections of the respiratory tract — especially for the elderly. Our results indicate, for example, that seasonal fluctuations were smaller during the last observed time period (1989–1998) for female as well as for male nonagenarians than for anyone during the period 1959–1968.

[16] The data were simulated as follows: $N_{50}^{\text{frail}} = 5 \times N_{50}^{\text{robust}}$; $q_x^{\text{frail}} = 0.06 + 0.0008 \times \text{age}$,

$$q_x^{\text{robust}} = \begin{cases} 0.06 + 0.0002 \times \text{age} & \text{, if age} \leq 87.5 \\ (0.06 + 0.0002 \times 87.5) + 0.0018 \times \text{age} & \text{, else.} \end{cases}$$

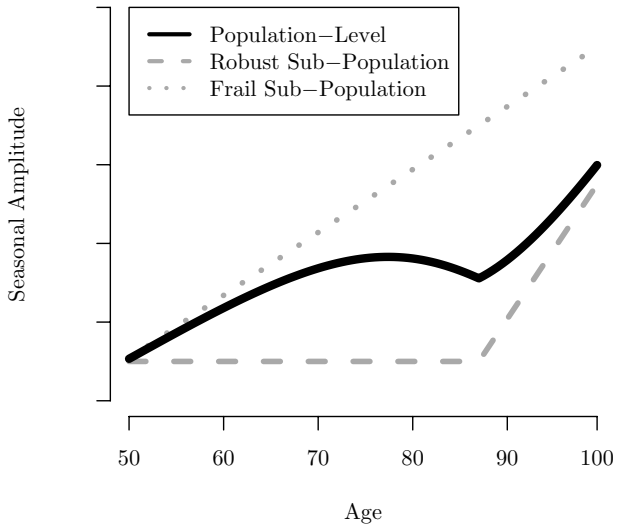



**Fig. 4.13.** Simulating the Impact of a Selection Effect on the Seasonal Amplitude

### 4.5.3 Seasonality by Region & Age

Figure 4.14 shows the development of seasonal mortality by age and region in the US for women and men for the last observed decade, 1989–1998. Because no reliable estimates turned out for Alaska and Hawaii, the two states have been omitted. No differences can actually be detected between the seven remaining regions. Also the possible mis-specification of some states from the group "Mountain" did not result in an estimation which differs from the other categories. One can see the aforementioned (cf. Figures 4.11 and 4.12) nonlinear increase of seasonality with age. All regions follow this pattern rather closely. These results are unexpected: Previous studies usually indicated that regions with a warm or moderate climate (e.g. the UK, Ireland, Portugal, Spain, Greece) tend to have higher seasonal fluctuations in mortality and deaths than colder regions such as Russia, Canada or Scandinavian countries [97, 98, 135, 147, 252]. This has usually been attributed to the fact that people in colder regions have higher indoor temperatures and avoid exposure to outdoor cold. If those findings could have been converted to the United States, one would assume that the regions "South Atlantic" and "East/West South Central" should show higher seasonality than other regions. According to the "Köppen Climate Classification", all states covered in these two regions belong to the "Humid Subtropical Climate". Surprisingly, they do not deviate in

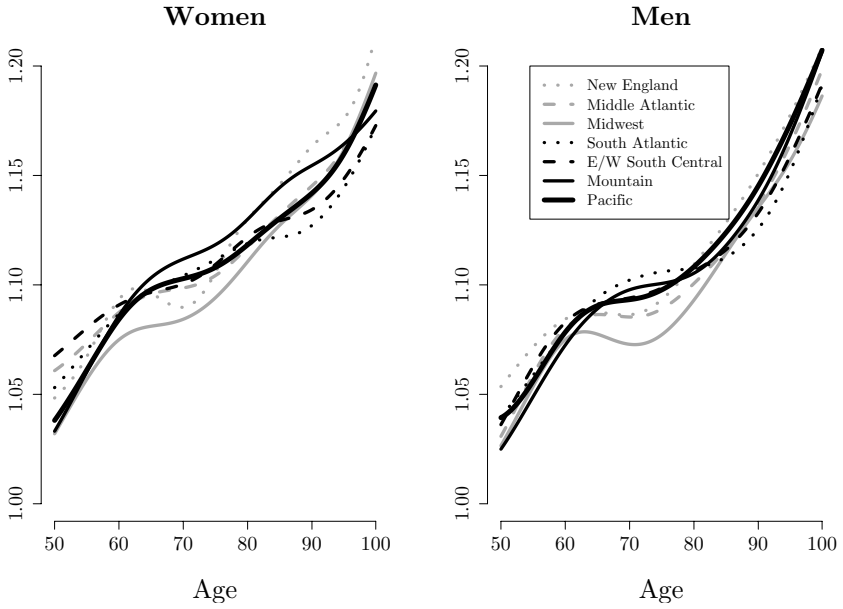



**Fig. 4.14.** Seasonality of All Cause Mortality by Age, Sex and Region, 1989–1998

any way from the other regions in the United States which are less humid and cooler. This underlines that social and cultural factors are important forces in shaping the seasonal pattern of deaths, as climate appears to be negligible. It has to be mentioned, though, that "region" in the United States is not only correlated with climate but also with socio-economic status and life expectancy. Residents in New England spent on average more time in school than women and men in the regions "South Atlantic" or "South Central".[17] At the same time, life expectancy is also lower in those regions [290]. This could suggest also an alternative explanation: there are two opposing forces which cancel each other out. On the one hand, the regional differences do exist as in Europe between warm and cold regions. That would imply that the southern states show higher seasonality than the states in the northeast. On the other hand, this differential is counteracted by a selection effect. Mortality is higher in the south of the United States. Due to these higher death rates, frail people tend to die at younger ages than in the North, which should have a rather depressing effect on seasonality. We consider the first explanation (no regional differences) to be more likely than the balanced outcome of two opposing forces. If the latter were true, it would require a social gradient by education: Due to a selection effect, people with low education should also

[17] Based on our own calculations using the number of years spent in school of deceased women and men. The results were similar for all ages above 50 as well as for people being 80 years old.

show lower amplitudes in their mortality fluctuations. As will be shown later in this chapter (page 118), a social gradient is observable — with the opposite direction, though: people with an academic degree have generally lower seasonality than people with only a few years spent in formal education.

### 4.5.4 Seasonality by Region & Period

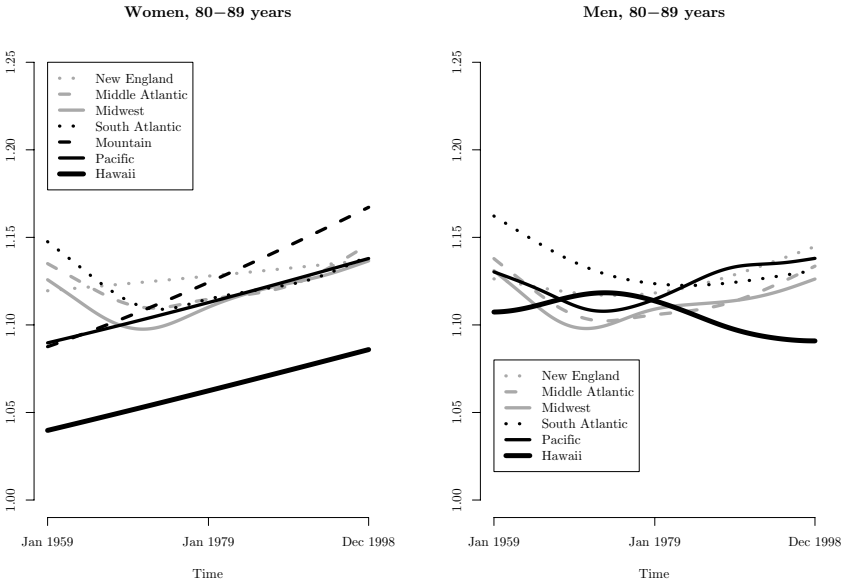



**Fig. 4.15.** Seasonality of All Cause Mortality by Region and Sex, 1959–1998

Figure 4.15 portrays how seasonal death fluctuations have changed over time in various regions of the United States. For reasons of clarity, only results for the age-group 80-89 years have been plotted. Due to numerical optimization problems,[18] it was possible to display only six regions for men (missing: "Alaska", "Mountain" and "East/West South Central") and seven for women (missing: "Alaska" and "East/West South Central"). Despite this unfortunate loss of information, several interesting features can be observed: The decrease in seasonality discovered in Figure 4.9 (page 108) did not occur in the US as a whole. Rather, three regions were responsible for this development for women and for men likewise: Middle Atlantic, South Atlantic and the Midwest. They

[18] While none of the $\lambda$-parameters reached one of their limits, no values for $\theta$ were possible to be input to lower the variance of the Pearson residuals anywhere close to 1.

showed decreasing seasonality for the first decade observed. All other regions already showed an increase during that period. With the exception of Hawaii (thick, solid, black line) trends have converged for the remaining regions since the late 1960s. This suggests that the existing climatic differences have become less and less relevant over time, as social circumstances and living conditions have become more alike in all regions. Hawaii represents an outlier — especially for women. One could either argue that seasonality in Hawaii is smaller than in other regions because of the predominant tropical climate. There, less precautions are required to avoid cold-related mortality during certain seasons as the temperature varies there less than in other (climatic) regions of the United States. It could also be, however, a statistical artifact due to the small number of deaths in Hawaii compared to the other analyzed regions. This latter hypothesis receives support from the study by Seto et al. [341] . They found differences of 22% between winter and summer mortality from coronary artery disease mortality . This shows that seasonal mortality in Hawaii does not differ from the United States as a whole, since we found roughly the same results in our description of winter/summer differences for cardiovascular diseases (Winter/Summer Ratio 1.206, cf. Table 4.2 on page 89).

### 4.5.5 Seasonality by Education & Age

Educational level serves as an indicator for socioeconomic status. How this variable affects seasonal fluctuations in deaths over age for women and men during the period 1989–98 is portrayed in Figure 4.16. For women and men alike, seasonal fluctuations are the highest for the category "not stated" given by the thin, dashed gray line. Apart from that residual category, a clear social gradient in seasonal mortality is observable until age 90. The biggest difference is to be seen between people who have earned a college degree or more (black solid line) and who have received no formal education at all (gray solid line). Persons who belong to the highest educational group have the lowest seasonal amplitude and vice versa. Again, it is remarkable how little women and men differ from each other in terms of seasonal fluctuations. The social gradient diminishes with age and vanishes completely for both sexes at about age 90. The path to convergence is interesting: People with highest completed education show a relatively steep slope whereas the pattern of people without any formal education is rather constant over time. One could therefore argue that people with relatively poor education face seasonal fluctuations in deaths throughout large parts of their adult lives which highly educated people only have to face at very advanced ages. Our estimates show that education does not matter for seasonal mortality when people are 90 years old. It is hard to make any inferences about the last years in our age span until the $100^{\text{th}}$ birthday. It seems as if people with the least formal education ("elementary school or less" depicted in the gray , solid line) do not become more susceptible to stressful environmental living conditions. Whether a direct effect or an

indirect (compositional) effect, or both, cause this stationary pattern is hard to answer. A direct effect would assume that people with low education are so weak in general that they die regardless of the current season. Contrastingly, a selection effect is also imaginable: As people with lower education tend to die at younger ages [374], only a highly selected subgroup is still alive at ages above 90. It is possible, that those people are especially strong in withstanding environmental stress during winter. This latter hypothesis receives further support when the development after age 90 is investigated for the other educational groups. A social gradient is still observable but the other way round. However, the ones facing higher seasonal fluctuations are highly educated people, whereas people with less education display smaller seasonal amplitudes. This pattern is possibly a reflection of a compositional effect as people with higher education are less selected than people with lower education.

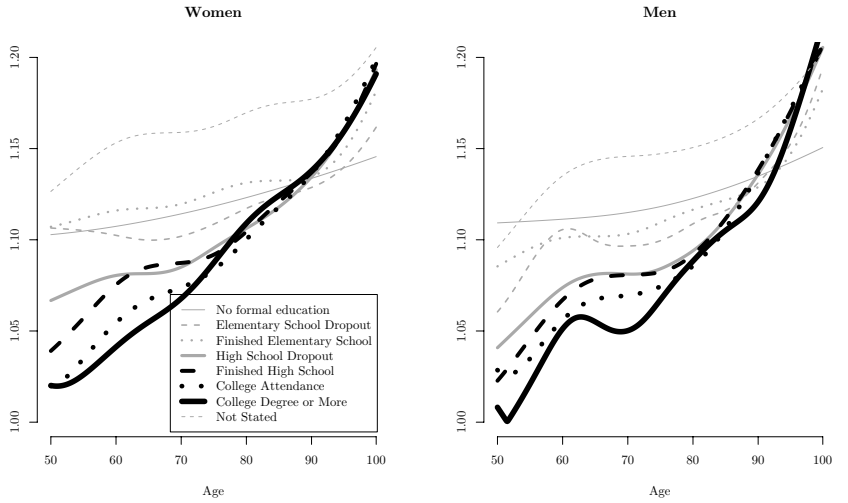



**Fig. 4.16.** Seasonality of All Cause Mortality by Age, Sex and Educational Status, 1989–1998

We investigated the influence of socio-economic status on seasonal mortality further by analyzing not only mortality from all causes but also from selected causes. The results for cardiovascular mortality are shown in the upper two panels of Figure 4.17, respiratory mortality is plotted in the lower two panels. Women's results are in the left column, men's seasonal fluctuations by age are displayed on the right. In all four panels we detect the aforementioned (Fig. 4.16) social gradient: The more years spent in formal education, the lower the seasonal fluctuations. One important difference is, though, that the relative differences for respiratory diseases are smaller than for cardiovascular diseases. Both causes of death are known to have a social

gradient [69, 210]. For mortality in general, however, the extent of the slope is larger for respiratory diseases than for cardiovascular diseases. This suggests that an inverse relationship of the social gradient exists across causes of death between general susceptibility and seasonal susceptibility.

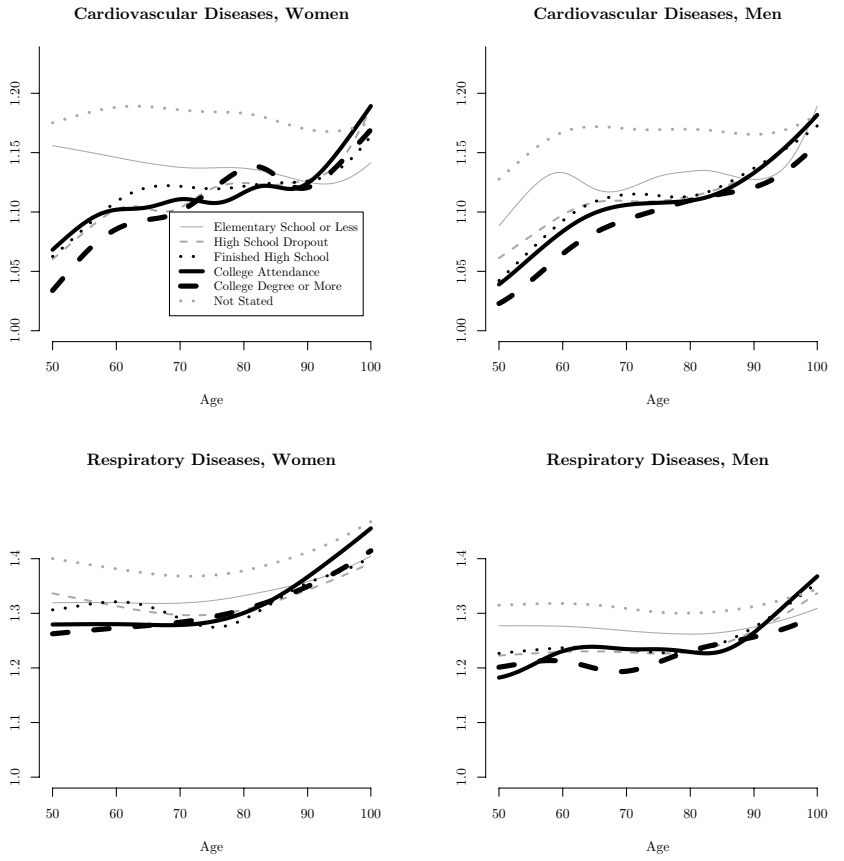



**Fig. 4.17.** Seasonality of Mortality from Cardiovascular and Respiratory Diseases by Sex and Educational Status, 1959–1998

### 4.5.6 Seasonality by Marital Status & Age

Seasonal differences in deaths by marital status are shown in Figure 4.18 by sex and age. Although the numbers of deaths by marital status vary considerably by marital status for women and men (cf. Table 4.4, page 92), the estimates for both sexes are again very similar. While the variable "education"

provided a clearly visible social gradient, "marital status" does not show such a clear-cut picture. Nevertheless, married people appear to have lower amplitudes in seasonal mortality across their life-course than widowed or never married people. This supports the idea of a protective effect of marriage also for seasonal mortality. Two possible causal pathways are: Married people can share their financial resources and are therefore able to have higher quality in housing and access to better medical care. It could also be the presence of another person in the household who is able to provide help in an emergency (e.g. calling an ambulance in case of a possible stroke). The lack of these factors is possibly reflected in the higher seasonality of never married and widowed people. Most likely these people live alone and don't have access to two sources of income. From mortality research in general it is known that divorced people are showing higher death rates than married people. In the case of seasonality, however, they are rather indistinguishable from married women and men. One could hypothesize for the US, therefore, that the presence of a partner is less important than the access to economic resources: divorced people are also likely to live alone. If this were decisive they should show similar seasonality as never married and widowed people. What makes them different is that they don't lose their financial resources.

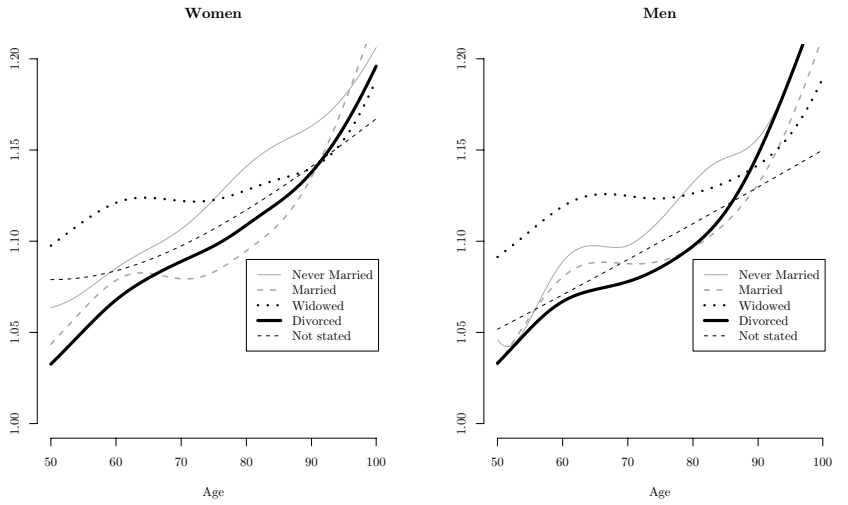



**Fig. 4.18.** Seasonality of All Cause Mortality by Age, Sex and Marital Status, 1989–1998

We note indications for a selection effect. At about age 90, the amplitudes in seasonality are converging among the analyzed marital status groups, suggesting even a crossover. This converging trend is also observed in studies on mortality in general [e.g. 125]. It can be argued that people who were

never married have typically a higher mortality rate throughout their life. Consequently, only a selected subgroup survives to those high ages, while the married are still more heterogenous in their composition with respect to frailty.

## 4.6 Summary

Seasonal fluctuations in deaths in the United States between 1959 and 1998 have been analyzed in this chapter. While models using information on the actual events (=deaths) and on the exposed population are preferable, sometimes only data on deaths are available — without any information on the individuals at risk. This analysis represents such an approach relying only on death counts. These data are derived from annual Public-Use-Files from the Centers for Disease Control and Prevention (CDC) in the United States. The time span covers the period 1959–98. Although deaths at all ages are included, our analysis restricts itself to the age-range 50–99 years. Almost 80 Mio. individuals died during that period in the given ages. They formed the basis of our analysis.

We developed a new method specifically designed to meet our needs in the presence of overdispersed count data. This analysis represents the first extensive application of this new method. We used a log-linear model where additive terms for the trend (one term) and for the season (at least two terms) were related to the mean of the observed deaths at a certain time or age via a log-link. These components are allowed to vary smoothly over time (or age). We fit this varying-coefficient model by using $P$-Splines which are the well-known $B$-Splines with a penalty on their respective regression coefficients. Thus, we did not impose any parametric form on either the trend or on the seasonal component but rather estimated changes over time (or age) data-driven. It has been shown with simulated data, that our new approach fits data with the given structure very well and much better than the standard methods.

Our analysis over calendar-time resulted in a slightly upward moving trend since the early 1970s for seasonal mortality from all causes as well as from cardiovascular and cerebrovascular diseases. This could reflect on the one hand that the differences between summer and winter mortality have become bigger on the individual level. The introduction of air conditioning and the widespread usage of central heating can serve as an explanation. It would imply that the former decreased summer mortality faster than the latter shrunk cold-related mortality. On the other hand, one can argue that compositional changes caused this increase over time. Because of the progress made in survival in general, relatively frail people attain high ages who would have died in the past at younger ages. They are most likely the ones who are more susceptible. In the case of respiratory disease we observed a decrease over time which could be attributed to the spread of central heating.

Seasonality of deaths is increasing with age. This increase is, however, neither linear nor monotonous. We observed, rather, a development in three stages: After an initial increase between ages 50 to 60/65, seasonality remains relatively constant for about twenty years after which they start increasing again. This puzzling pattern — especially for cerebrovascular diseases — may hint at an interaction between "real" changes in susceptibility (=increasing trend) and compositional changes due to mortality selection (=depressing effect).

In European countries large variations in seasonality have been observed between countries with warm, moderate, and cold climate. This pattern has not been reflected in our regional analysis of the US. The examination by age showed the expected trajectory of an increase as people are getting older. Nevertheless, the slope does not differ if people are living in a rather warm or cold state. Our analysis over period shows a converging trend over time which is probably caused by a tendency towards similar social circumstances and living conditions in all regions of the United States.

Seasonality in deaths by educational status has not been investigated previously. Our decomposition approach resulted in a clear social gradient. The lower the educational status, the larger are the differences between winter and summer. This effect can be observed until about age 90 when all educational groups display more or less the same seasonality. Beyond age 90, we observed a crossover which might have been caused by a selection effect: while frail people with low education are most likely already dead, frail people with a college degree are still alive and are more likely to die in winter than the rather healthy, homogeneous group with lower education.

Our explorative approach into the question whether marital status is as important for seasonal mortality as for mortality in general was not as successful as the investigation into educational status. Married women and men appear to have lowest seasonal fluctuations over age, while never married and widowed people have higher seasonality. Unfortunately, the trajectories of the four analyzed marital status groups are partly overlapping. This implies that a straightforward distinction as for mortality in general is not possible.

This analysis of seasonality in deaths in the United States found support for the surprising finding of previous studies of increasing seasonality over time. Cardiovascular and cerebrovascular diseases follow this trend rather closely, whereas respiratory diseases showed a decreasing trend. Our "three-stage-increase" of seasonality with age showed that a statement like "seasonality increases with age" is too simple. The most important findings from our study are:

- We found no differences in seasonality by region — neither over time nor by age — as we could have expected from previous literature on Europe. This underlines the importance of social factors compared to climate.
- In a pilot approach of analyzing the importance of education on seasonal mortality, we detected a strong social gradient. The higher the educational

status, the lower is the seasonal fluctuation in death for most of the adult life.

- The most important finding is probably the lack of differences of seasonal fluctuations in seasonal mortality for women and men. While women face throughout their entire life course lower mortality than men, the relative differences between winter and summer seem to be negligible.