3

Measuring Seasonality

3.1 Introduction

Public health measures aim to improve the health of the people. For that purpose, it is an absolute necessity to discover the origins of diseases. If diseases, and ultimately mortality, occur seasonally, "an environmental factor has to be considered in the etiology of that disease" [244, p. 275]. An enormous diversity of causes of death has been related to seasonal incidence: cardiovascular diseases [420], asthma [40], infectious diseases [260], diarrhea and cholera [31, 391], suicide [139], and congenital malformations [90, 184] to name only a few.

The aim of this chapter is to present the methods that have been suggested and/or employed in the literature and to discuss their advantages and disadvantages by using hypothetical and real data. From a methodological point of view, one can basically distinguish between two categories of studies. On the one hand, studies that test for the existence of seasonal trends and, on the other hand, studies that examine whether certain covariates are correlated with seasonal fluctuations in mortality. The latter group has already been briefly presented in [139]. A thorough discussion of all methods is not the scope of the present study: it is almost unfeasible to inspect all methods such as correlation analysis, regression analysis (linear, logistic, Poisson, . . .), analysis of variance (ANOVA), etc., which have been employed for studies of seasonality.

This chapter is only concerned with the first group, i.e. statistical approaches to detect, measure and test seasonality. Thus, we remained in a univariate framework by not including any covariates apart from time or age. Within the methods analyzed, we can make a further distinction into three subdivisions:

- *Indices* to Measure the Extent of Seasonality
- *Statistical Tests* for Seasonality
- *Time-series Methods* for Seasonality

The rationale behind these methods will be introduced and is followed by a discussion of the their respective pros and cons. The three groups will then be faced with hypothetical and real data to evaluate how sensitive they are to various sample sizes and different distributions. The last part of this chapter will summarize the findings and give recommendations which method should be applied in which situation.

The presented and evaluated time-series methods have already been implemented by various statistical computer packages. Apart from one test (χ^2-Goodness-of-Fit test), no ready-to-use software was available for any of the indices or tests. Therefore, these indices and tests have been implemented in the R-language [170, 301]. The actual code can be obtained from the author.

3.2 Seasonality Indices

3.2.1 Introduction

Most researchers did not perform any statistical test to analyze if a seasonal pattern is present in a population or not. Instead, they used some descriptive tools to characterize the pattern they found in their data. The simplest representations are monthly death counts. This method was especially widespread among scientists of the 19th century, as they did not have any sophisticated methods or computers at their disposal. Tulloch's analysis, for example, examined the seasonality in mortality among the British Troops in the West Indies by revealing monthly death counts [368].

However, even some early researchers used some descriptive tools that are still common nowadays. In 1912, Lucien March calculated an index for which he standardized the annual number of deaths to 1,000 [240]. Thus, values above $83\frac{1}{3}$ indicated above average mortality; values below $83\frac{1}{3}$ stood for mortality less than what could be expected from a uniform distribution of deaths across the twelve months. Many recent studies used by and large the same standardization. But instead of a radix of 1,000, the preferred choice is 1,200. Thus, the expected number for each month in a uniform distribution is 100 which makes it more apprehensible for users of the decimal system to detect above- and below-par mortality.

For example, the "Cambridge Group for the History of Population and Social Structure" used this index in their explorations of English population history [415, 416]. Studies on contemporary mortality also use this "100-Index" [e.g. 101] which is easy to calculate and interpret.

Besides writing a table with the number of counts or the values of monthly mortality rates, there are also other possibilities to make the seasonal distribution of deaths comparable over time and/or across populations. The easiest way is a barplot with 12 categories representing the months on the x-axis and the usage of bars or lines to represent the actual monthly values (see Figure 3.1 as an example).

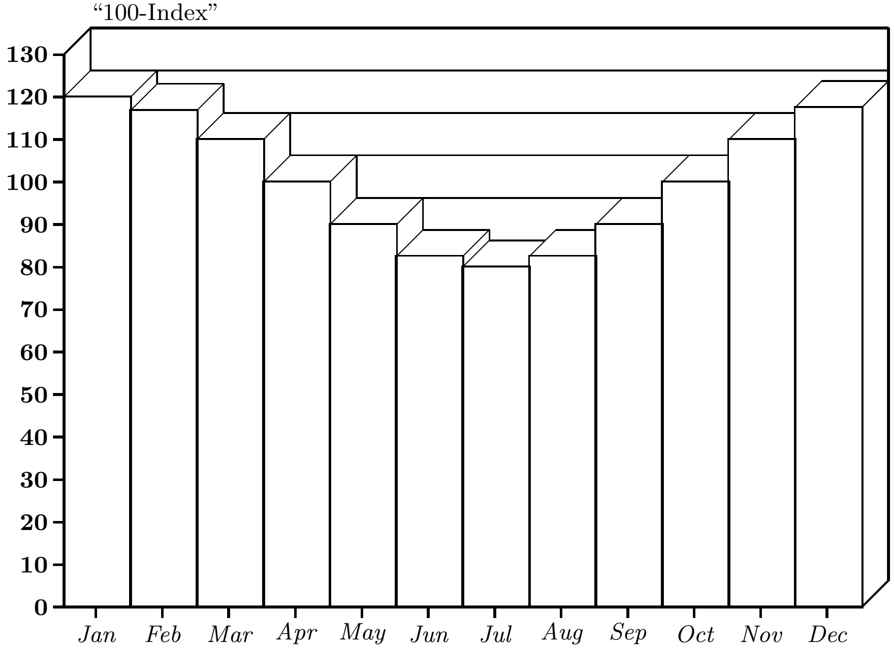

Fig. 3.1. Graphical Representation of Seasonal Mortality Data (Hypothetical Data)

3.2.2 Winter/Summer Ratio

However, a mere graphical description fails to satisfy a researcher as the judgment in comparing two populations (or one across time) depends largely on eyesight. Thus, statisticians have employed countless indices to describe data with one number (e.g. the median as measure of central tendency for an ordinal variable). An uncomplicated index for seasonality is a mortality ratio where winter mortality is divided either by summer mortality or by the average mortality during the year. With the index φ_1 in Equation 3.1, we opted to divide the number of deaths in winter by the number of deaths in summer.

$$\varphi_1 = \frac{\sum\limits_{i=JAN}^{MAR} Deaths_i}{\sum\limits_{j=JUL}^{SEP} Deaths_j} \tag{3.1}$$

Such an index has several desired properties. For example, it is easy to interpret. "1" would indicate that there is no difference between summer and winter deaths. Values above one correspond to more winter than summer deaths (and vice versa). A value of 1.24 would indicate that the number of deaths is 24 percent higher in winter than in summer. Thus, it gives a measurement of the differential between winter and summer deaths but does not take into account what happens in other months. In addition, the choice for the basis of the numerator and the denominator is somehow arbitrary.

3.2.3 Concentration/Dissimilarity Indices

Most other seasonality indices can be interpreted as a measurement of concentration or of dissimilarity. Two central concepts in that area are the Lorenz-Curve and the Gini-Coefficient. The construction and the interpretation of the Lorenz-Curve is straightforward. Assume we have a population with a certain characteristic, e.g. income (which is the typical example in textbooks). The first step is to order the population by this characteristic and give each individual a rank. For each rank, one calculates the proportion of all people whose rank is smaller or equal to that rank. Simultaneously, you also compute for each rank the relative frequency of income earned by people whose rank is smaller or equal to the specific rank [4]. If you plot these two cumulative relative frequencies, the result will be a Lorenz-Curve, as shown in Figure 3.2. If the variable of interest is uniformly distributed, the result would be the solid black curve connecting the points $(0,0)$ and $(1,1)$ with a straight line. If the variable of interest is unequally distributed, the curve still starts at $(0,0)$ and ends at $(1,1)$. But it will bend and, by definition (because of the sorting procedure), must be convex to the x-axis [192] as shown by the dotted line in black in Figure 3.2.

Several indices try to express the degree of inequality in a certain population based on the Lorenz-Curve. Among them, the Gini-Coefficient is the "best known and most widely used measure of divergence [...]. It is defined as an area between the diagonal and the Lorenz Curve, divided by the whole area below the diagonal" [346, p. 310]. Despite its intuitive appeal, the Gini-Coefficient has some important drawbacks for analyzing seasonality in deaths: it is defined for continuous data. Our data, however, are usually given in discrete units i.e. months. This shortcoming is not too problematic. It has been shown before for other discrete data, that the Gini-coefficient can be adapted to this situation [e.g. 346]. More important is the following dilemma:

- Either the monthly values are ordered according to their rank as intended by this procedure. It would then be almost certain that the original order of the months is not preserved and we could only answer the question whether our data deviate from a uniform distribution or not. We cannot make any claims about the shape of the deviation.
- The other approach one could follow is not ordering the data (i.e. the first category is January, the second category is February, ...). In that case we cannot exclude the possibility that the Lorenz-Curve is crossing the diagonal and the Gini-Coefficient would be not defined for that situation. The solid gray line in Figure 3.2 indicates this: A typical seasonal distribution in an unordered way would not only cover an area in the lower right triangle for which the Gini-coefficient is defined.

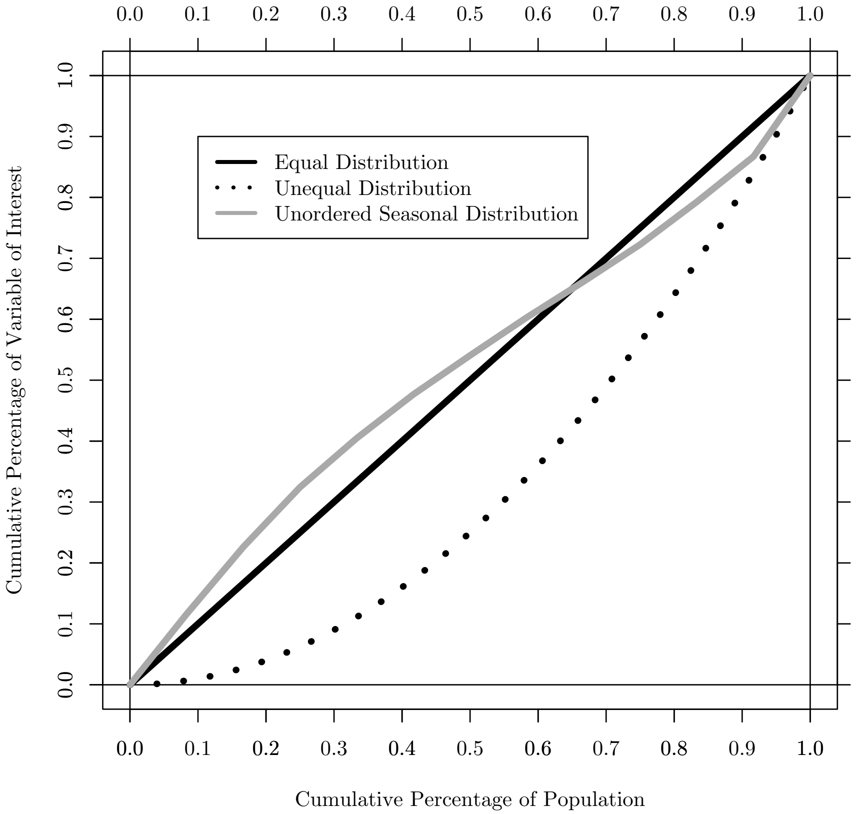

Fig. 3.2. The Lorenz Curve — Hypothetical Examples

In his study of seasonal mortality in Sweden, John Wilmoth [407] did not use the real Gini-Coefficient but a related measurement (φ_2) derived from the analysis of residential segregation [403]:

$$\varphi_2 = \frac{1}{2}\sum_{i=1}^{12}|p_i - q_i|, \tag{3.2}$$

where p_i is the *observed* proportion of deaths in month i, q_i is the *expected* proportion of deaths in month i. $\sum_{i=1}^{12} p_i = \sum_{i=1}^{12} q_i = 1$; in our case of a uniform hypothetical distribution $q_1 = q_2 = \ldots = q_{12} = \frac{1}{12}$.

As long as we use relative frequencies and a uniform hypothetical distribution, the value of φ_2 ranges from 0 in the case of equal counts in all months to 0.91666 ($= \frac{1}{2}\left(11 \times \left|0 - \frac{1}{12}\right| + 1 \times \left|1 - \frac{1}{12}\right|\right)$) in the case when events only occur in one month. Although this approach seems to be fruitful at first sight, it has a major disadvantage. For real data, the value of φ_2 does not exceed 0.1. Most emipirical distributions of deaths have a value around 0.03. Another drawback is its insensitivity to the ordering of the months. It does not take into account if a peak is followed by a trough by a peak by a trough, etc. or if there is only one peak and one trough.

Closely related to dissimilarity indices are measurements of concentration. They can also serve as an index for seasonality. The best known is entropy. This concept has been developed in information technology and measures the degree of uncertainty. It was introduced to demography in the mid 1970s by Lloyd Demetrius [65]. In popular terms, entropy tells you how safe a guess is when you do not know anything about the exact distribution of the variable of interest. In the case of a uniform distribution, your guess would be very unsafe as each category would be equally probable. Entropy, in this case, would reach its maximum value. If one uses a standardized index, entropy would be 1. If the distribution is getting closer to a monopolistic situation, entropy approaches zero. A relative entropy index (φ_3) with a defined maximum of 1 serves as our seasonality index [392, p. 22f]:

$$\varphi_3 = \frac{H(A)}{H(A)_{\max}} = \frac{\log_2(n) - \frac{1}{n}\sum_{i=1}^{12} n_i \log_2(n_i)}{\log_2 k} = \frac{\log_2(n) - \frac{1}{n}\sum_{i=1}^{12} n_i \log_2(n_i)}{\log_2 12}, \tag{3.3}$$

where n_i is the number of events in month i and $\sum_{i=1}^{k(=12)} n_i = n$; $\log_2$ is the *logarithmus dualis*, the logarithm to the base 2.

3.3 Tests for Seasonality

Besides these descriptive measurements, several statistics have been proposed to test for seasonality. They can be broken down into three groups: the χ^2-Goodness-of-Fit test and the “Kolmogorov-Smirnov-Type-Statistic” belong both to the group of *Goodness-of-Fit-Tests*; harmonic analyses based on Edwards' contribution [84] are members of the *Edwards' Family*. The third group consists of *Nonparametric Tests*.

3.3.1 Goodness-of-Fit-Tests

The χ^2-Goodness-of-Fit Test

The χ^2-Goodness-of-Fit Test is relatively popular for detecting seasonality because of its simple mathematical theory, which makes it easy to calculate and understand [139]. Pearson introduced the concept in 1900 [286] which can be applied to a variety of statistical problems [20]. Generally speaking, this test can be employed whenever the research question is: "In the underlying population represented by a sample are the observed cell frequencies different from the expected cell frequencies?" [344, p. 95] Thus, we test whether our empirical data can be a sample of a certain distribution with sampling error as the only source of variability [256]. Usually, this hypothetical distribution is a uniform distribution. However, there is no restriction on the underlying distribution. This test requires a sample from a population with an unknown distribution function $F(x)$ and a certain theoretical distribution function $F_0(x)$. The χ^2-Goodness-of-Fit Test examines the Null-Hypothesis $H_0 : F(x) = F_0(x)$ against the alternative hypothesis $H_A : F(x) \neq F_0(x)$. The test-statistic T is calculated as follows:

$$T = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right] \tag{3.4}$$

where $i = 1, \ldots, k$ are the groups in the sample. For seasonality studies, the value of k is usually 12. O_i and E_i are the observed and expected cell frequencies of the i^{th} class, respectively. If $F_0(x)$ is a uniform distribution, then $E_1 = E_2 = \ldots = E_k$.

T is under H_0 asymptotically (for $n \to \infty$) χ^2-distributed with $\nu = k - 1$ degrees of freedom [158, 321]. The χ^2-Goodness-of-Fit Test has been recently used, for instance, for the analysis of seasonality in suicide, myocardial infarction, diarrhea, pneumonia and overall mortality [110, 149, 207, 308, 345, 369, 391]. The major problem of the test is that the value of T is not asymptotically χ^2 distributed for small sample sizes. "In this case, the χ^2 statistic has positive bias, that is, it tends to be larger than the theoretical chi-square value it is supposed to estimate" [158, p. 239]. Various rules of thumb have been proposed for when the approximation is acceptable.[1]

The typical data on seasonality do not violate any of these restrictions of the use of the χ^2-Goodness-of-Fit Test. For seasonality studies, usually $\nu = 11$ and more than the suggested 5, 10, etc. cell frequencies are observed. In addition, the result of this test does not depend on the starting point (e.g. January, February, or any other month) as does the following test in its original version [278].

[1] For instance:

- E_i has to be ≥ 5 for each cell [344].
- Only if $\nu \geq 8$ and $n = \sum_{i=1}^{k} O_i \geq 40$ it is allowed to have expected frequencies of 1 in some classes [321].
- $k > 2$ and $n\pi_i^0 \geq 10$ for all i [392].

A Kolmogorov-Smirnov-Type-Statistic

The original Kolmogorov-Smirnov-Goodness-of-Fit Test (KS-Test) is comparable to the χ^2-Goodness-of-Fit Test in several ways. Both approaches are designed to test if a sample drawn from a population fits a specified distribution. In addition, the tests are not restricted to a certain class of distributions. Unlike the χ^2-Goodness-of-Fit Test, the KS-Test does not compare observed and expected frequencies for single classes, but rather the cumulative distribution functions between the ordered observed and expected values. This test was introduced in 1933 by Kolmogoroff.[2] Six years later Smirnoff provided a more elementary proof of it [204].[3] Generally speaking, the KS-Goodness-of-Fit Test has greater power than the χ^2-Goodness-of-Fit test and "is especially useful with small samples" [354, p. 708]. As for the χ^2-Goodness-of-Fit test, the Null-Hypothesis $H_0 : F(x) = F_0(x)$ for all $x \in \mathbb{R}$ is tested against the Alternative Hypothesis $H_A : F(x) \neq F_0(x)$ for at least one $x \in \mathbb{R}$. However, the ordinary Kolmogorov-Smirnov test contains some disadvantages. The first problem we face is that this test relies on ungrouped data from continuous distributions [393]. Also the modified method by Kuiper in 1962 is no longer valid "once the values [...] have been grouped into months" [113]. Another problem is the choice of the starting point. Although January is usually taken, it is somehow arbitrary. But — as pointed out in several articles — the result and its interpretation depends on the starting point [e.g. 250]. If one has to choose between (the described Pearson's) Goodness-of-Fit χ^2-test and the ordinary KS-Test, Slakter advises to use the χ^2-test as it is more valid than the Kolmogorov Test — even for small sample sizes and a uniform hypothetical distribution [351]. Freedman proposed an improved version, which eradicates both drawbacks: the problem of the starting point and of the grouping of data [113]. The hypothetical cumulative distribution (in our case a uniform distribution) is denoted by $F(t) = \frac{t}{12}$, where t equals the rank of each month of the year (January=1, February=2, ..., December=12). The sample cumulative distribution is denoted by $F_N(t) = \frac{j}{N}$, where j is the number of events (e.g. deaths) that have happened during all months $\leq t$. The test-statistic T is [113, 305]:

$$T = V_N\sqrt{N} = \sqrt{N}\left[\max_{1\leq t\leq 12}(F_N(t) - F(t)) + \left|\min_{1\leq t\leq 12}(F_N(t) - F(t))\right|\right] . \quad (3.5)$$

The distribution of T does not follow any specified distribution (e.g. χ^2; $N(\mu, \sigma^2)$, ...). Therefore this distribution has been empirically determined by performing Monte Carlo simulations and is tabulated in Freedman's article

[2] Spelling of Russian names (especially -ov vs. -off) differs not only in this dissertation but also in the original papers. Therefore, I opted to use the spelling in each case from the respective source document.

[3] In this article, Kolmogoroff refers to the articles [203] and [352].

[113]. Freedman's modified KS-Type Test has been used for the study of birth seasonality [e.g. 390].

3.3.2 Edwards' Family

The first statistical test especially designed for seasonality — or more generally speaking for cyclic trends — is Edwards' Test published in 1961 [84]. It is "the most cited and the benchmark against which other tests are evaluated" [394, p. 817]. Several others modified this test in order to be valid for small sample sizes or to allow for a different alternative hypothesis. These extensions will be presented after the discussion of the original contribution. All of them use sine- and cosine-waves to approximate the observed pattern, and therefore they are methods which belong to *harmonic analysis* [142, p. 641].

Edwards' Test

The underlying idea of the original test [84] is relatively straightforward and based on a geometrical framework [263]. Given a circle whose circumference is divided into k equally long parts. In the case of months per year, $k = 12$. Thus, each month's contribution to the surface of the circle is a sector of 30 degrees: January from 0° to 30°, February from 30° to 60°, ... and finally December from 330° to 360°. This is shown in Figure 3.3 (page 48).

A weight, N_i, is attached to the center of each segment (i.e. for January at 15°, for February at 45°, ...). N_i is the number of events in month i. If events were uniformly distributed, the center of gravity of this "wheel" would be the geometrical center of the whole circle as indicated by the small black circle in Figure 3.3. If, however, there is a considerable "pulse" or an underlying sinusoidal pattern, the center of gravity shifts away from the geometrical center. The small gray circle could be an example of a concentration of events in winter and, more precisely, the middle of January. If one is testing such a cyclical hypothesis against a uniform distribution, Edwards' Test has a higher power than the χ^2-Test [358]. Walter and Elwood extended Edwards' approach by allowing unequal expected numbers in each category [396]. In its original version, the Null-Hypothesis assumes to have equally spaced sectors with the same frequencies in each division. The allegation that the assumption of twelve equally spaced time intervals may cause problems in practice [139] can easily be refuted. One simply has to standardize the number of incidences according to the specific length of month. The test statistic $T = \frac{1}{2}a^2N$ is calculated as shown in Equation 3.6 (multi-line notation of T is taken from the original article [84]):

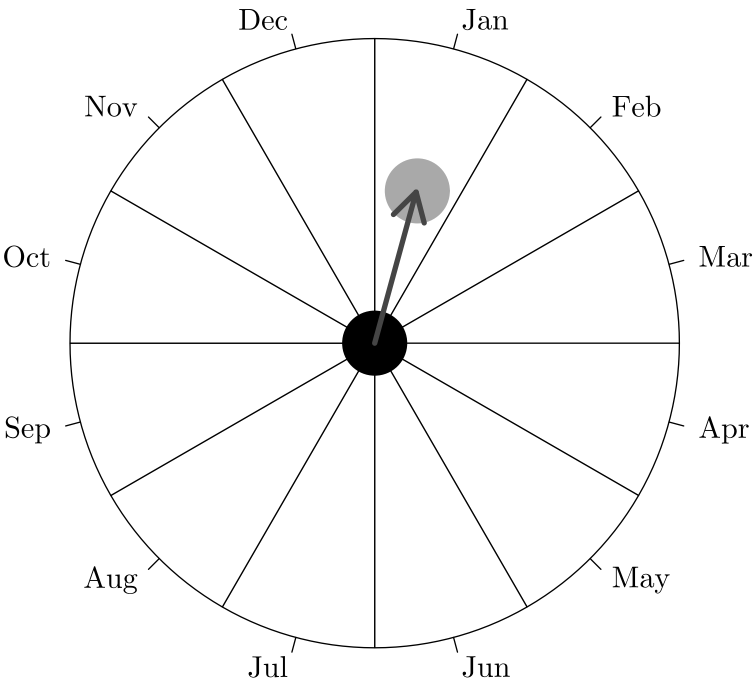

Fig. 3.3. Graphical Representation of Edwards' Test for Seasonality

$$\begin{gathered}S = \sum \sqrt{N_i} \sin \Theta_i \\ C = \sum \sqrt{N_i} \cos \Theta_i \\ W = \sum \sqrt{N_i} \\ d = \frac{\sqrt{(S^2 + C^2)}}{W} \\ a = 4d\end{gathered} \tag{3.6}$$

N_i corresponds to the number of events (e.g. deaths) in month i and $\sum_{i=1}^{k(=12)} N_i = N$. The parameter Θ_i indicates the position of the weight of each month on the wheel. Thus, Θ_i equals 15°for January, 45°for February, ..., and 345°for December. $\frac{1}{2}a^2 N$ is under H_0 asymptotically χ^2-distributed with two degrees of freedom [84]. Edwards' method has been employed in the study of coronary heart disease [340], myocardial infarction [131], and overall mortality [148, 268].

Roger's Test

As pointed out by Roger [312], Edwards' test does not yield satisfactory results for small and medium-sized samples. "This has the effect of making the type I errors in the test too large and hence leading to too many spurious significant results" [312, p. 153]. Roger tried to tackle the shortcoming of Edwards' Test for small sample sizes and proposed the following test statistic T [312]:

$$T = \frac{2\left[\left\{\sum_{i=1}^{k(=12)} N_i \sin\left(\frac{2\pi i}{k}\right)\right\}^2 + \left\{\sum_{i=1}^{k(=12)} N_i \cos\left(\frac{2\pi i}{k}\right)\right\}^2\right]}{n} \tag{3.7}$$

N_i represents the number of events in month i, and $n = \sum_{i=1}^{12} N_i$. T is under H_0 approximately χ^2-distributed with $n = 2$ degrees of freedom [244]. According to Roger, his test and Edwards' original test are equivalent for large samples. Roger's extension provided a useful tool for the analysis of "seasonal variations in variceal bleeding mortality and hospitalization in France" [30].

Pocock's Method

Pocock's [291] analysis of seasonal variations in sickness absence belongs also to the group of tests using harmonic analysis like [84] and [312]. While Rogerson [312] extended Edward's approach for small sample sizes, Pocock relaxed the assumption of a sinusoidal underlying pattern and "allows the alternative hypothesis of a seasonal pattern of arbitrary shape" [139, p. 49].The test statistic T, which was originally designed for weekly values for spells of sickness absence, has been slightly adapted for our approach of monthly death counts. It tests "the seasonal sum of squares" for deviations from the Null-Hypothesis that deaths occur randomly in time.

$$T = k \sum_{j=1}^{\frac{k}{2}} \frac{\left(a_j^2 + b_j^2\right)}{2\overline{A}}, \tag{3.8}$$

where k represents the number of intervals (12 months) and $j = 1, 2, \ldots, \frac{k}{2}$. $\overline{A}$ is the mean of the monthly number of deaths A_i in month $i = 1, 2, ..., 12 (= k)$.

$$\left.\begin{aligned} a_j &= \frac{2}{k}\sum_{i=1}^{k} A_i \cos\frac{2\pi ij}{k} \\ b_j &= \frac{2}{k}\sum_{i=1}^{k} A_i \sin\frac{2\pi ij}{k} \end{aligned}\right\} \qquad j = 1, \ldots, \frac{k}{2}$$

T is under H_0 approximately χ^2 distributed with $\nu = 11$ degrees of freedeom.

Cave and Freedman's Method

Cave and Freedman proposed another modification of Edwards' test [44]. Instead of a sinusoidal curve with only one peak and one trough per year, they allowed two maxima and two minima. Their test-statistic is, thus, relatively similar to [84]. The difference is the implementation of the Θ-parameter: while for Edwards' test [84] it is required to calculate $\sin\left(\frac{2\pi\Theta_i}{360}\right)$,[4] one proceeds for the method of Cave and Freedman [44] by computing $\sin\left(\frac{2\pi\Theta_i}{180}\right)$, thus standardizing 2π to 180°.

3.3.3 Nonparametric Tests

Hewitt's Test and Rogerson's Extension

Edwards [84] mentioned that his test was only one approach to measure seasonality. He explicitly considers also a nonparametric alternative whose construction is relatively similar to a simple Run-Test [393]. Based on that brief suggestion — two paragraphs in Edwards' original article — Hewitt et al. elaborated a nonparametric test based on rank-sums [150]. While Edwards suggested "to consider the ranking order of the events which are above or below the median number" [84, p. 83], Hewitt et al. [150] propose to use "all the ranking information rather than a simple dichotomy" [150, p. 175]. According to them, the monthly frequencies are ranked. The month with most occurrences (e.g. deaths) will have the value "12" assigned. Consequently, "1" indicates the month having the least events. Keeping the original order of the months (e.g. starting with January and ending with December), we can calculate the rank-sums of six consecutive months (January–June, February–July, . . . , December–May). The test statistic T is the maximum value that one of the rank-sums attains. T can range from 21(=1+2+3+4+5+6) to 57(=12+11+10+9+8+7) and is symmetrically distributed. The authors suggest referring to the upper tail of the cumulative distribution for significance testing which they tabulated in their article based on 5,000 Monte-Carlo trials. Not surprisingly, their empirical results correspond closely to Walter's exact significance levels for Hewitt's test calculated nine years later [395]. Using such a test based on ranks has the advantage that one "avoids the problem of specifying a particular algebraic version" [113, p. 225] of what is meant by seasonal fluctuation. However, it lacks power for small and moderate sample sizes [113]. Besides, this test cannot be applied — as Reijneveld [305] points out — if there are ties. For our analysis of mortality with relatively large samples, though, ties seem to be quite unlikely. Of more relevance are the objections of Rogerson [315], Wallenstein [394] and Marrero [244] to "the assumption that the year is split into two equally wide intervals of 6 months each" [315, p. 644]. While Wallenstein and Marrero take a "one-pulse model" also into

[4] this applies obviously also to $\cos\left(\frac{2\pi\Theta_i}{360}\right)$

consideration, Rogerson develops a generalization of Hewitt's Test for peak periods of 3-, 4-, and 5-months [315]. Similar to taking the maximum rank sum of all possible combinations of six consecutive months, Rogerson uses the maximum rank sum of any consecutive three, four, or five month period, respectively. Because of its relative simplicity to calculate, Hewitt's Test has enjoyed widespread use [315]. However for the analysis of seasonal mortality it has not been employed as often as Edwards' Test or the χ^2-Goodness-of-Fit Test. To my knowledge, Akslen's and Hartveit's application to seasonal variation in melanoma deaths has been the only application of it so far [1]. Apart from Walter's exact specification of significance levels for Hewitt's test [395], the distributions of the respective test statistics were based on relatively few randomly generated sequences of data. The appendix (Section B.1, page 181) shows results from my own simulations.

David-Newell-Test

Another nonparametric alternative was proposed by David and Newell [64]. Their suggestions, however, have not received much attention. In contrast to Hewitt's non-parametric test for seasonality, one does not use the ranking information but the actual number of events.

$$T = \max_i \left| \frac{M_i - M_{i+6}}{\sqrt{N}} \right| \tag{3.9}$$

where N_j is the number of events in month j and $M = \sum_{j}^{j+5} N_j$; $N = \sum_{j=1}^{12} N_j$. The test statistic T does not follow any standard distribution. Therefore the critical values for two significance levels ($\alpha_{0.01}, \alpha_{0.05}$) are given in their paper [64].

3.4 Time-Series Methods

3.4.1 Introduction

The previous sections have focused on indices and statistical tests to describe seasonality and test for seasonality in data grouped into one year. Contrastingly, the following sections deal with the analysis of seasonal time-series. Typically, these data are either count data or rates over time.

Most analyses of seasonal time-series data have the opposite aim than our approach: conventionally, researchers try to "seasonally adjust" the time-series. This means that one wanted to get rid of the seasonal "distortions" to identify the "true effect". We, however, are interested in seasonality itself: How does the seasonal pattern change over time? Despite these two antagonistic theoretical starting points, the actual analyses can be carried out with the same methods because both approaches need to model the exact seasonal signal from the data.

Basically, there are two approaches for seasonal time-series-analysis: either one decomposes the time-series into several components, or one models all of these aspects simultaneously [389]. In reality, methods for analyzing seasonal time-series cannot always be clearly assigned to one of the groups as they are using methodology from both strains. In the following paragraphs, I want to briefly outline what is meant by decomposition methods and simultaneous modelling. Subsequently, I will discuss several of the methods which are actually used and also implemented in various software packages.

3.4.2 Decomposition Methods

It is argued that decomposing time-series started in the 1920s at the National Bureau of Economic Research (NBER) [417] of the United States. Starting with the first "monthly means method" and the "ratio to moving average method" [270] to modern methods, decomposition methods are based on the assumption that the observed data contain four components [335]:

Trend: The trend is the long-term change in the time-series. In the analysis of seasonal mortality two thrusts can be imagined to influence the trend over time: First, a change in the variable of interest: death rates are falling rapidly for people above age 70 at least for the last 30 years [378]. Secondly, a compositional change can either increase the effect of the variable of interest or it can be counteracted. The latter is more probable for the analysis of death counts as more and more people attain very high ages because of improved survival conditions [383].

Cycle: The cyclic component captures a fluctuation with a frequency of more than one year [335]. While they are an important part of economic analysis, e.g. the Kondratieff long economic cycles [205], they play only minor role in mortality research.[5] The cyclic component is sometimes not extracted on its own but rather as a part of the trend component.

Season: The seasonal component is an annually repeating pattern observed in the time-series, and is the feature of the data which is our main focus. While it is beyond doubt that climate shapes the basic pattern of seasonal mortality fluctuations, a large body of literature shows that the impact of climate can be mediated and alleviated. Consequently, we want to analyze how seasonal fluctuations are changing over time, which measures indirectly the influence of improvements in public health and general living conditions.

Irregular: The remainder between the aforementioned components and the observed data is summarized in the irregular component.

[5] I consider the analysis of Stoupel et al. [359] concerning the impact of "space proton flux" on the temporal distribution of cardiovascular deaths as negligible.

Basically, there are two approaches on how these components constitute the observed time-series: In an additive model, one assumes that the trend component y_t^t (includes the cyclic component), the seasonal component y_t^s and the irregular component y_t^{res} are working independently. Thus, the resulting model would be:

$$y_t = y_t^t + y_t^s + y_t^{res} \tag{3.10}$$

In the majority of real-world applications, however, independent effects are rather the exception rather than the rule. Thus, a multiplicative combination of the trend and the seasonal components

$$y_t = y_t^t \times y_t^s + y_t^{res} \tag{3.11}$$

is often preferable.

3.4.3 Simultaneous Modelling

In contrast to the decompositon approach, the time-series data can also be modelled simultaneously. This is done by so-called seasonal ARIMA-Models. This approach follows the Box-Jenkins methodology [32] of identifying parsimonious models for the data under scrutiny. A seasonal ARIMA-Model is an extended ARIMA-Model. An ARIMA-Model is an extended ARMA-Model. Thus, I want to start with the basic model: An ARMA-Model consists of an autoregressive (AR) and of a moving average (MA) part. As explained in [95], "an $AR(p)$ process is specified by a weighted average of past observations going back p periods, together with a random disturbance in the current period [...], an $MA(q)$ process is specified by a weighted average of past random disturbances going back q periods, together with a random disturbance in the current period."[6] The aim of ARMA modeling is a parsimonious model. This means in the words of its creators to "employ the smallest possible number of parameters for adequate representations" [32, p. 16]. Typical diagnostics to check for pickung the best model are, for example, the Akaike Information Criterion (AIC) or the Schwarz Bayesian Criterion. However, ARMA Modelling requires a stationary time-series. If the data are non-stationary which is rather the rule than the exception, the ARMA(p, q) Model is extended to an ARIMA(p, d, q) Model (ARIMA=Auto Regressive Integrated Moving Average). In such an ARIMA-Model, the time-series is first differenced finite d times until a stationary process is obtained. Seasonal ARIMA-Models represent a further extension. The general form of such a SARIMA Model is:

$$\text{ARIMA}(p, d, q)(P, D, Q)_{12}.$$

[6] Mathematically, the specifications may be written as given by Box et al. [32, p. 52]:

$$\begin{aligned}\text{AR(p)} : \tilde{z}_t &= \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \ldots + \phi_p \tilde{z}_{t-p} + a_t \\ \text{MA(q)} : \tilde{z}_t &= a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}\end{aligned}$$

In addition to the previously explained parameters p, d, q, SARIMA contains the parameters P, D, Q indicating autoregressive (P) and moving average (Q) components differenced D times at a seasonal lag. In the case of an annual seasonal pattern with monthly values, the respective lag is 12 months.

3.4.4 Seasonal Time-Series Methods

The "Classical" Decomposition Method

The "classical" decomposition method uses moving averages as outlined in Brockwell and Davis [34] or Hartung [142]. The first step is an estimation of the trend "by applying a moving average filter specially chosen to eliminate the seasonal component and to dampen the noise"[34, p. 30]. The seasonal component is then estimated by computing the average deviation of the monthly values from the estimated trend. This method is, however, irrelevant for the rest of this chapter, as it contains a constant seasonal component. The aim of this research is, though, exactly the analysis of this seasonal component over time (or age).

X-11

Still the most widely used method is the so-called "X-11, Census II" method. Its development can be traced back to the "ratio to moving average method" from the 1920s. The various revisions have been labeled "X-" followed by the version number. X-11 was developed at the U.S. Bureau of the Census in 1965 by Julius Shishkin [417].

The estimation is performed in several steps [cf. 417]. Ghysels and Osborn [122] and Yaffee [417] give an overview how these steps are performed. We are following the overview given by Fischer [108] for the X-11 ARIMA variant:[7]

1. First estimate of the seasonal and the irregular component using a 12 term moving average.
2. Preliminary estimate of the seasonal factors using a 5-term moving average.
3. A 12-term moving average is applied to the preliminary factors found in the previous step.
4. The seasonal factor estimates are divided by the seasonal irregular ratio to obtain an estimate of the irregular component.
5. Detection of outliers
6. Adjustment for the beginning and the end of the time-series (necessary since symmetric filters are used).

[7] X-11 uses moving averages for the estimates. Since these weights are symmetric, problems arise in the beginning and in the end of the time-series. To remedy this drawback, Statistics Canada introduced the so-called X-11-ARIMA/88 to improve the fore- and back-casting possibilities of X-11 [cf. 417, 55–56].

7. Estimation of preliminary seasonal factors by applying a weighted 5-term moving average to the SI (ratio of the seasonal and irregular component) ratios with replacement of extreme values (detected two steps earlier).
8. Step 3 is repeated and applied to the factors in step 7.
9. Division of the original data by the result from the previous step to obtain a preliminary seasonally adjusted time-series.
10. The original series is divided by the result of applying a moving average to the seasonally adjusted series.
11. Applying a weighted 7-term moving average to each month's SI ratio separately. This results in a second estimate of the seasonal component.
12. Step 3 is repeated.
13. The original series is divided by the result from step 11 to obtain a seasonally adjusted time-series.

Fischer [108, p. 15] gives a flow-chart to display graphically this procedure. Despite its popularity, several serious drawbacks of X-11 have been pointed out [14, 53, 303]:

- Using X-11 can imply that a non-seasonal cycle can be wrongly specified as seasonal.
- X-11 is not very robust in the case of a sudden change in the trend. This might sound unimportant as natural processes typically do not change all of a sudden. However, in the analysis of seasonal mortality of a specific cause of death across time, relatively abrupt changes in the trend can happen after an ICD-Revision[8] — no matter how careful the preparation of the time-series.
- In the case of zero-value observations ($\neq$missing values), neither an additive nor a multiplicative X-11 approach is applicable. Zero events might happen in certain age-groups for diseases with a highly seasonal pattern like deaths from influenza.
- X-11 may over- or under-estimate the seasonal component (non-idempotency). The lack of this property is a serious shortcoming for the analysis of seasonal changes over time.
- The values for the seasonal factors depend on the beginning of the time-series. As pointed out by Raveh [303], X-11 yields different seasonal estimates for the same original values if the series is shifted forward for half a year, for instance.
- X-11 is over-sensitive to outliers. This problem of the original variant, though, seems to be eradicated by X-12.

Despite these disadvantages, X-11 is still a popular choice. The original version has been employed for the analysis of seasonal mortality in an early 20th century population [21] and more recently for examining seasonal deaths in the United States [102]. An example for X-11-ARIMA is Richard Trudeau's

[8] ICD is the abbreviation for "International Statistical Classification of Diseases". See http://www.who.int/whosis/icd10/ .

study on "monthly and daily patterns of death" in Canada [367]. For our analysis, we used X-12-ARIMA which is the successor to X-11-ARIMA "to handle additive outliers and level shifts" [35, p. 1]. The estimates should be at least as good as X-11 since it improves the detection and correction of outliers and estimates automatically the ARIMA-Models [108, p. 16].

SABL

William S. Cleveland and his colleagues did not only pin-point the weaknesses of X-11, they also suggested alternative procedures. Their first suggestion was the so-called SABL [50, 51].[9]. This procedure works basically in four steps [51]:

1. A power transformation of your time-series
   The power transformation is carried out as follows [52, p. 53]:

$$x^{(p)} = \begin{cases} x^p & \text{if } p > 0 \\ \log_e x & \text{if } p = 0 \\ -x^p & \text{if } p < 0 \end{cases} \tag{3.12}$$

   The value of the power p should be chosen to minimize the interaction between the trend and the seasonal component. Fortunately, the program Splus picks the best p-value — provided a vector of possible values has been given before.
2. Additive decomposition of the transformed time-series into trend, seasonal, and irregular component. The details of this decomposition are described in [p. 15–16 51]:[10]
   a) "A combination of smoothers, which involve moving medians for robustness, is used to get initial estimates of the trend and the seasonal. Moving medians are similar to moving averages except that means are replaced by medians.
   b) The irregular, which is the series minus the trend and seasonal, is computed.
   c) Robustness weights are computed using the irregular values. Irregular values large in absolute value receive small or zero weight.
   d) Updated estimates of the trend and seasonal are computed using smoothers that are doubly-weighted moving averages. The two sets of weights are those computed in step (c) and the usual kind of weights in moving averages.
   e) Steps (b) to (d) are repeated using the updated estimates of trend and seasonal. The trend and seasonal component in step (d) on the second pass are the final trend and seasonal."

[9] SABL is the abbreviation for **S**easonal **A**djustment at **B**ell **L**aboratories

[10] Alternatively, one can also consult the flowchart given in [108, p. 12].

3. Seasonal Adjustment. This step is not required in our estimations since we are interested in the final seasonal component obtained in the previous step.
4. In its original version, SABL printed tables and plotted graphs. We do not need this option since we are using modern statistical software which allows for printing the results and plotting the graphs in a user-defined way.

STL

The seasonal decomposition STL was also invented by Cleveland and his colleagues [48]. As usual, the data are decomposed into three components: a trend, a seasonal part and a remainder. Among other criteria, the authors wanted to develop a procedure which has a simple design, its use is straightforward, does not have problems with missing values, has a robust trend and seasonal component, and is easily and quickly implemented on a computer. The core of the procedure are smoothing operations based on locally-weighted regression (loess). As written by Cleveland et al. [48, p. 6]: "STL consists of two recursive procedures: an inner loop nested inside an outer loop. In each of the passes through the inner loop, the seasonal and trend components are updated once; [...] Each pass of the outer loop consists of the inner loop followed by a computation of robustness weights; these weights are used in the next run of the inner loop to reduce the influence of transient, aberrant behavior in the trend and seasonal components." The inner loop consists of the following steps [48, p. 7–8]:

1. Detrending
2. Smoothing of the Cycle Subseries
3. Filtering of the Smoothed Cycle-Subseries (obtained from pervious step)
4. De-trending of Smoothed Cycle-Subseries
5. De-seasonalizing
6. Trend Smoothing

BV4

BV4 is the abbreviation of the fourth revision of the so-called "Berliner Verfahren". It is the official seasonal adjustment method of the Statistisches Bundesamt (Federal Statistical Office) in Germany. It was developed by Martin Nourney and is described in detail in [275, 276, 277]. Currently the Statistisches Bundesamt is replacing BV4 with an updated version called BV4.1 which can handle calendar effects and outliers better. According to Speth [357] three of the main advantageous characteristics of BV4.1 are:

- Low cost benefit ratio because high-quality analysis can be performed without expert knowledge and without much experience for time-series decomposition methods.

- Results are independent from the user.
- High efficiency of the seasonal adjustment which can incorporate even rapid changes in the seasonal component.

BV4.1 assumes (after a possible transformation of the data, e.g. a log-transform) an additive decomposition of the time-series of the form [104, 105]:

$$Y_t = G(t) + S(t) + \epsilon_t. \tag{3.13}$$

$G(t)$ denotes the trend-cycle component ("Glatte Komponente") which is approximated by a third-order polynomial: $G(t) = \hat{y}_t^t = a_0 + a_1 t + a_2 t^2 + a_3 t^3$. The seasonal component $S(t)$ is approximated by 11 trigonometric functions [see also 108]:

$$y_t^s = \sum_{i=1}^{5} (b_i \cos \lambda_i t + c_i \sin \lambda_i t) + b_6 \cos \lambda_6 t.$$

The irregular component ϵ_t is an independently, identically distributed random variable with mean 0 and a constant variance σ^2. The actual fitting procedure is performed by locally weighted least squares.

TRAMO/SEATS

TRAMO/SEATS has been developed by Victor Gómez and Agustín Maravall.[11] Their work is based on "seasonal adjustment by signal extraction" by Burman [39]. A detailed, technical description of TRAMO/SEATS is given in [239]. Fischer [108] summarizes the six steps of the TRAMO/SEATS procedure as follows (a flow-chart with more details is given on page 18 of [108]):

- TRAMO identifies automatically an ARIMA Model
- Simultaneously, outliers are detected
- TRAMO passes its results to SEATS
- "In SEATS, first the spectral density function of the estimated model is decomposed into the spectral density function of the unobserved components, which are assumed to be orthogonal" [108, p. 17]
- Then, the trend-cycle and the seasonal component are estimated
- In the last step, outliers are re-introduced.

3.5 Evaluation of Seasonality Indices and Tests Using Hypothetical and Real Data

3.5.1 Description of Datasets

We distinguish between *real* and *hypothetical* data. If real data were taken from publications which introduced a measurement for seasonality, the data

[11] TRAMO stands for **T**ime Series **R**egression with **A**RIMA Noise, **M**issing values and **O**utliers. SEATS stands for **S**ignal **E**xtraction in **A**RIMA **T**ime **S**eries.

were used to check the correct implementation of the underlying algorithm. Otherwise real data from other sources were used to analyze how the various indices and tests behave in typical situations of seasonal mortality analyses. They are briefly described in the next section and are plotted in Figure 3.4. Hypothetical data served only experimental purposes. For example, we think that any measurement should test positively if a pronounced sine wave is present. Random numbers and a uniform distribution should, conversely, not detect seasonality at all. For the hypothetical data, I have always created two data-sets: one with a small sample size and one with a larger sample size as described below. For each category, only the data based on the larger sample is plotted in Figure 3.5.

Real Data

Wrigley: These data consist of 75,398 deaths from the British parish register data between 1580–1837. These data represent standardized death counts where 100 indicates the mean number of monthly deaths. Wrigley et al. [415] provide more details.

Nuns and Monks: Marc Luy kindly provided death counts from his data collection on Bavarian nunneries and monasteries [229]. A detailed description can be found in [228]. The nuns' data-set consists of 3,919 individuals who have died during the 20th century in the analyzed nunneries. In the other data-set all 349 male deaths are included which occured in the respective monasteries during the 19th century.

Union Army: These data are taken from the *Public Use Tape on the Aging of the Veterans of the Union Army* [111]. It consists of 24,610 individuals who died between January 1862 and December 1937. Each death has been recorded by month and year of death. Thus, the aggregated data-set contains 912 records. For our analysis we only used deaths starting in 1866 to avoid distorting effects due to the Civil War.

Danish Register Data: All Danes are included who were alive on 1 April 1968 and 50 years or older and died by August 1998. The data from which these 1,176,383 deaths have been derived are explained in more detail in [72].

Respiratory Diseases: The data-set includes 25,272.56 men who have between January 1959 and December 1998 from respiratory diseases in the United States being between 80 (inclusive) and 90 (exclusive) years of age. The reason for the non-integer number of deaths is that monthly deaths have already been adjusted to the same length. The data were taken from the public use files of the *Centers for Disease Control and Intervention (CDC).* These data are described in more detail in Chapter 4.

Anencephalics: The monthly distribution of 176 cases of anencephalics that have occurred in Birmingham between 1940 and 1947 are given in [84].

Lymphoma: The monthly distribution of 133 cases of Burkitt's lymphoma from the West Nile district of Uganda between 1966 and 1973 are given in [113].

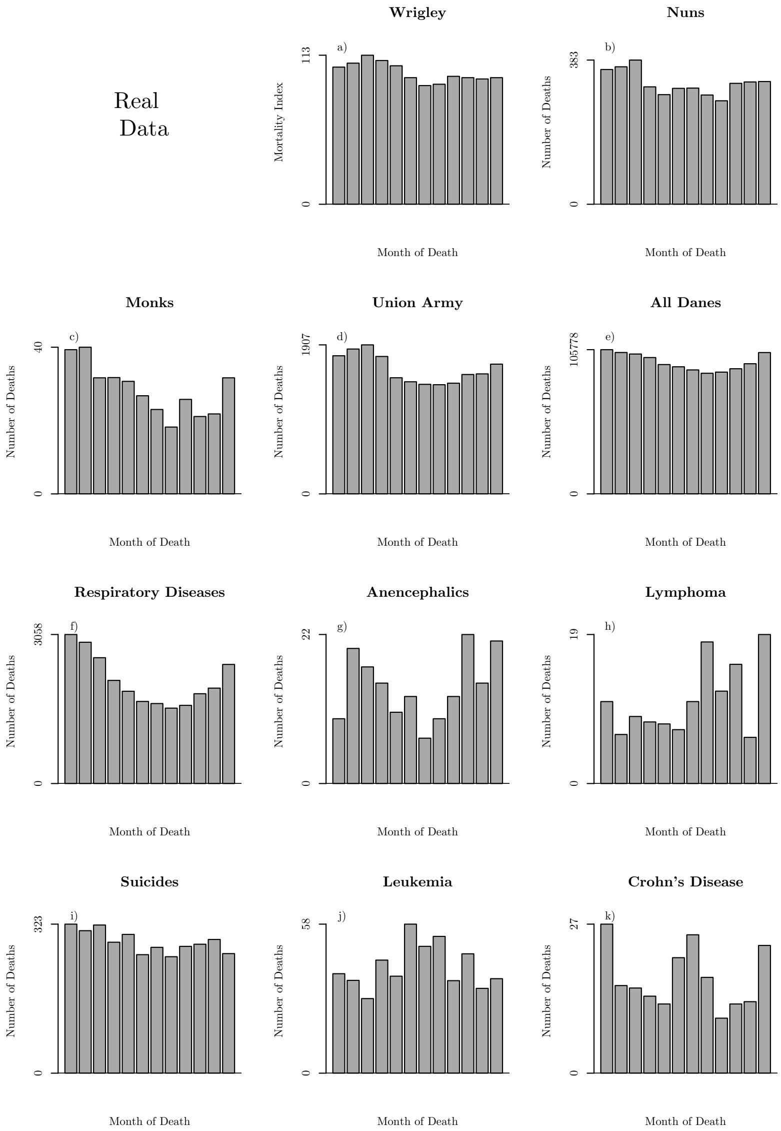

Fig. 3.4. Graphical Representation of Real Data-Sets

Suicides: The monthly distribution of adolescent suicides (3474 cases) in the United States, 1978–79, is given in [315].

Leukemia: The monthly distribution of the onset of acute lymphatic leukemia between 1946 and 1960 (506 cases) is taken from the *British National Cancer Registration Scheme* as reported in [64].

Crohn's Disease: Cave and Freedman [44] give a bar-plot displaying the monthly distribution of the onset of Crohn's disease for 211 patients in three British hospitals between 1945 and 1974. More details about the data can be found in the original article.

Hypothetical Data

Uniform Distribution: Two vectors with 12 elements — each of the 12 elements representing the numbers of death in a month —are given consisting of either 5 or 5000 cases in each month.

Sine Wave: Again, we have two vectors with 1 entry for each month. The "small" sine wave has a maximum value of 12 in January and 8 in July, whereas the large sample's extreme values are 120 and 80 in the same months.

Cosine Wave: The Cosine Waves are equivalent to the two Sine Waves with a forward shift of $\frac{3}{2}\pi$. One should expect the same results as for the Sine Wave data as we basically face the same pattern. Testing the measurements with the Cosine data can help to evaluate whether certain indices or tests are restricted to the Northern Hemisphere with a peak in the first (few) months of the year.

Local Summer Peak: The literature on seasonal mortality sometimes also refers to a second peak in summer. The data-sets are the same as the Sine Wave data apart from the minimum. Instead of values of 8 and 80 respectively in July, we have values of 10 and 100.

One-Pulse Pattern: Some causes of death do not have a sinusoidal but a "one-pulse"-pattern. This means that deaths are uniformly distributed throughout the year with the exception of some months where deaths rise rapidly. Our data have 10 (small sample) and 100 (large sample) deaths in each month. In winter, however, deaths suddenly increase, reaching a peak in January and February of 13 and 130 deaths, respectively.

Random Pattern: Randomly distributed numbers should (in general) not result in significant test results for seasonality. The random numbers are derived from the "true" random number generator at `http://www.random.org`. Integers were generated between 900 and 1100 for the larger sample; for a smaller sample we used the same numbers but divided each of them by 10.

3.5.2 Results and Discussion for Indices and Tests

Results and Discussion for Indices

Table 3.1 shows the results for the three descriptive indices $\varphi_1, \varphi_2, \varphi_3$. The upper section refers to hypothetical data, in the lower section we faced the indices with real data. In our synthetically generated data only one value is

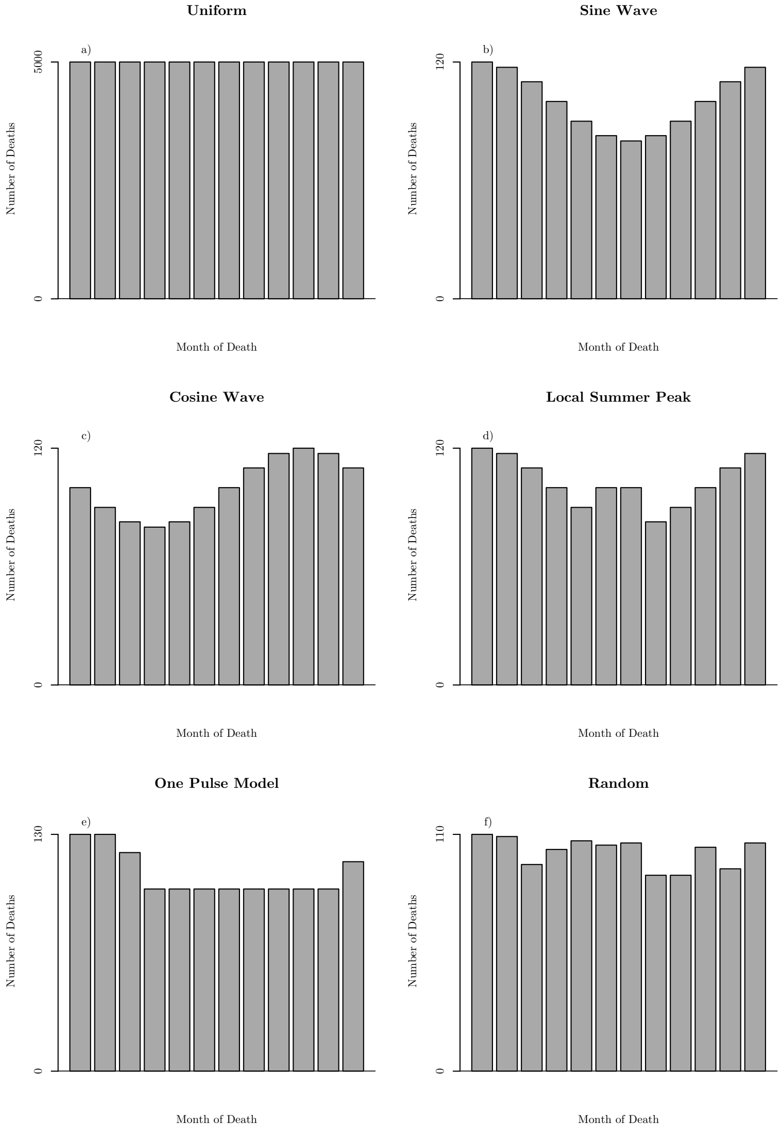

Fig. 3.5. Graphical Representation of Hypothetical Data-Sets

given for each pattern (uniform distribution, sine wave, . . .) as all indices are inelastic with regard to sample size.

As mentioned in their description above, seasonality indices are closely related to measures of inequality. Goodwin and Vaupel [126] suggested several

Table 3.1. Sample Sizes and Results for Descriptive Indices φ_1 (Winter/Summer), φ_2 (Dissimilarity), and φ_3 (Entropy) for Seasonality of Hypothetical and Real Data

Hypothetical Data	Sample Size N	φ_1	φ_2	φ_3
Uniform	60 / 60,000	1.000	0.000	1.000
Sine Wave	120 / 1,200	1.375	0.062	0.996
Cosine Wave	120 / 1,200	0.833	0.062	0.996
Local Summer Peak	123.73 / 12,373.32	1.274	0.048	0.997
One-Pulse Pattern	129.5 / 1,295	1.267	0.049	0.998
Random Pattern	122.2 / 1,222	1.094	0.029	0.999

Real Data	Sample Size N	φ_1	φ_2	φ_3
Wrigley	1,199	1.165	0.032	0.999
Nuns†	3,919	1.266	0.038	0.998
Monks†	349	1.656	0.100	0.989
Union Army Veterans†	24,610	1.191	0.034	0.999
Danish Register Data†	1,176,383	1.161	0.028	0.999
Respiratory Diseases	25,272.56	1.781	0.102	0.989
Anencephalics†	176	1.605	0.139	0.978
Lymphoma†	133	0.627	0.167	0.969
Suicides†	3,474	1.191	0.034	0.999
Leukemia†	506	0.749	0.087	0.992
Crohn's Disease†	211	1.113	0.132	0.982

† Monthly values have been adjusted to equal weights

desirable properties for "Measures of Evenness". Most of them can also be applied to the field of seasonality:[12]

The Relativity Principle: The relativity principle refers to sample size. According to the principle, it is a desirable property for any index that it should be independent from the sample size, as long as the proportions of each category remain the same. An index which fulfills this condition will return the identical value for a data-set of 100 individuals and 100 million individuals if the corresponding subgroups in each population contribute the same share. All our three indices fulfill this condition. To produce the upper part of Table 3.1, we used data with the same basic distribution and only varied the sample size. When we analyzed the data, φ_1, φ_2 as well as φ_3 gave exactly the same results.

The Transfer Principle: According to this principle, a diversity measure should increase if there are any transfers from a "poor" individual to a "rich" individual. Applied to the case of seasonality in mortality, any good index

[12] The so-called *Anonymity Principle*, for instance, is not included. This principles states that an index should be anonymous in the sense that it does not matter which element of the underlying population has a certain trait. It is less useful for the analysis of seasonality because a seasonality index should actually take into account whether January or September shows higher mortality values.

should increase if there is a transfer from a low mortality month such as June to a high mortality month like December.[13] All indices fulfill this property as well. For instance, if a certain number of deaths occur less in summer but more in winter, the winter/summer ratio φ_1 would increase, likewise φ_2, the dissimilarity index. As entropy φ_3 is measuring concentration it decreases, consequently. It has to be mentioned, though, that not all possible transfers affect φ_1. If there are any transfers between spring and autumn months, this winter/summer ratio will remain constant.

Standardization: Standardizing an index to a certain interval, say $[0; 1]$, facilitates describing population across time or across countries. The dissimilarity index φ_2 fulfills this condition. In the case of a uniform distribution, its value is 0. If deaths occur only in one month, it reaches its maximum value (for the case of 12 possible event times) 0.91666. In the same scenario, entropy (φ_3) would be bounded by 1 (uniform distribution = "minimum safeness of a guess") and would approach 0 in the case where deaths are only possible in one month. The winter/summer index is only bounded on one side. If death is equally probable in each month, φ_1 would be 1. If deaths only occurred in summer, φ_1 would approach 0; on the other extreme if deaths exclusively happened in winter, $\varphi_1 \rightarrow \infty$.

Intelligibility: "Ideally, a measure should be easy to comprehend, intuitively meaningful, simple to explain to others, and naturally relevant to the problems addressed" [126, p. 11]. The winter/summer ratio is the only index which fulfills all these conditions, especially the explanation to other people of the dissimilarity index or of entropy is considerably more complicated for φ_2 and φ_3 than for φ_1. It also seems to be more meaningful intuitively. For example, a value of 1.26 from φ_1 (real data: nuns) can be read as: among nuns in the respective data-set, 26% more died during winter than during summer. The corresponding values of $\varphi_2 = 0.038$ and $\varphi_3 = 0.998$ can contribute only little to the understanding of the underlying phenomenon.

Based on these criteria, it is difficult to make a decision for which index is best suited for seasonality studies. It can be argued that the winter/summer ratio φ_1 is preferable because of its better intelligibility and because φ_2 and φ_3 are unfavorable due to the following reasons:

- Standardized entropy (φ_3) does not seem to be a useful index because we observed only values between 0.996 and 1 for hypothetical data and between 0.969 and 0.999 for real data in our analysis. As this index is standardized to have a value range of $(0; 1]$, φ_3 uses only roughly 3% of its potential range. The dissimilarity index φ_2 performs only slightly better than φ_3 in that respect (18% of the value range is used).

[13] This, of course, holds only for measurements of unevenness. If we measure concentration, the opposite direction should be true.

- Neither index can distinguish between two patterns where one has its peak in winter and the other one has its maximum value in summer (i.e. φ_2 and φ_3 would give the same results in both situations).
- A related problem is the order of the months: The indices φ_2 and φ_3 do not take the ordering of the months into account: It does not matter, for instance, for φ_2 or for φ_3 whether the values appear as in a sine wave or in any other order. Clearly, an unfavorable property of any index for seasonality.

Results and Discussion for Seasonality Tests

Figures 3.6 and 3.7 (pages 66, 68) show the results of our analysis of the tests described in section 3.3. The tests are ordered according to which group they belong to: Goodness-of-Fit tests, the "Edwards' family" or nonparametric tests. All of them are faced with the data-sets outlined in section 3.5.1. Hypothetical as well as real data were tested for two levels of significance: $\alpha_1 = 0.95$; $\alpha_2 = 0.99$. In the case of hypothetical data, we tested both sample sizes as indicated by "small" and "large". To facilitate recognizing the outcomes of these tests, they were labeled with a dark gray square and a "$-$"-sign in case of insignificant results at the given level. A light gray square and a "$+$"-sign were used for significant values.

All tests passed a minimum requirement: as displayed in Figure 3.6, none of the tests detects seasonality for a uniform distribution nor for the random pattern — regardless of the sample size. The tests developed by Cave and Freedman [44] and by Pocock [291] will be excluded from further analysis, as they did not evaluate any of the hypothetical data-sets to be seasonal [44] or only the sine/cosine-data based on a large sample [291].

An advantage of all tests presented here is that they show exactly the same results for a sine and a cosine wave if the sample size is the same in both instances. This implies that all of them can be applied on both hemispheres giving the same results. While this requirement sounds obvious, the most widely-used seasonal time-series method, X-11, does not produce the same results if data start in January or in June [303]. Nevertheless, it is quite surprising that neither any Goodness-of-Fit-test nor any test from the "Edwards' Family" tests positively for seasonality for the sine- and cosine curves when the sample size is small. Only the non-parametric tests yield significant values. Because of their definition (using ranks instead of the actual counts or rates), Hewitt's tests and its generalization by Rogerson output the same values for small and large sample sizes. For the data-sets with a local summer peak or displaying only one pulse, we again detect the sample-size dependency for the Goodness-of-Fit tests and for the "Edwards' Family": no seasonality for small samples, significant ρ-values for large samples. The nonparametric tests for peak periods of 3 and 4 months behave as expected by returning significant results for the hypothetical data with one-pulse.

Hypothetical Data (Part A)	Uniform				Sine Curve				Cosine Curve			
	small		large		small		large		small		large	
Test	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
1. Goodness-of-Fit-Tests												
1.1 Chi-Square - Goodness of Fit	-	-	-	-	-	-	+	-	-	-	+	-
1.2 Kolmogorov-Smirnov-Type-Statistic	-	-	-	-	-	-	+	+	-	-	+	+
2. Edwards' Family												
2.1 Edwards' Test	-	-	-	-	-	-	-	-	-	-	-	-
2.2 Roger's Extension of Edwards' Test	-	-	-	-	-	-	-	-	-	-	-	-
2.3 Pocock's Method	-	-	-	-	-	-	+	-	-	-	+	-
2.4 Cave and Freedman	-	-	-	-	-	-	-	-	-	-	-	-
3. Non-Parametric-Tests												
3.1 Hewitt's Test [1]	-	-	-	-	+	-	+	-	+	-	+	-
3.2 Rogersons' Generalization for												
3.2.1 a 5-months peak [1]	-	-	-	-	+	+	+	+	+	+	+	+
3.2.2 a 4-months peak [1]	-	n.a.	-	n.a.	+	n.a.	+	n.a.	+	n.a.	+	n.a.
3.2.3 a 3-months peak [1]	-	n.a.	-	n.a.	+	n.a.	+	n.a.	+	n.a.	+	n.a.
3.3 David-Newell-Test	-	-	-	-	-	-	+	+	-	-	+	+

Hypothetical Data (Part B)	Local Summer Peak				One Pulse Pattern				Random Pattern			
	small		large		small		large		small		large	
Test	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
1. Goodness-of-Fit-Tests												
1.1 Chi-Square - Goodness of Fit	-	-	-	-	-	-	-	-	-	-	-	-
1.2 Kolmogorov-Smirnov-Type-Statistic	-	-	+	+	-	-	+	+	-	-	-	-
2. Edwards' Family												
2.1 Edwards' Test	-	-	+	+	-	-	+	+	-	-	-	-
2.2 Roger's Extension of Edwards' Test	-	-	+	+	-	-	+	+	-	-	-	-
2.3 Pocock's Method	-	-	-	-	-	-	-	-	-	-	-	-
2.4 Cave and Freedman	-	-	-	-	-	-	-	-	-	-	-	-
3. Non-Parametric-Tests												
3.1 Hewitt's Test [1]	+	-	+	-	-	-	-	-	-	-	-	-
3.2 Rogersons' Generalization for												
3.2.1 a 5-months peak [1]	+	+	+	+	-	-	-	-	-	-	-	-
3.2.2 a 4-months peak [1]	+	n.a.	+	n.a.	+	n.a.	+	n.a.	-	n.a.	-	n.a.
3.2.3 a 3-months peak [1]	+	n.a.	+	n.a.	+	n.a.	+	n.a.	-	n.a.	-	n.a.
3.3 David-Newell-Test	-	-	+	+	-	-	+	-	-	-	-	-

1) The actual levels of significance for the non-parametric tests are:		
Levels of Significance	0.05	0.01
Hewitt	0.0483	0.0130
Rogerson 5 months peak	0.0562	0.0152
Rogerson 4 months peak	0.0470	n.a. (Max. Ranksum=42; p_{42}=0.0267)
Rogerson 3 months peak	0.0545	n.a. (Max. Ranksum=33; p_{33}=0.0545)

Fig. 3.6. Results for Seasonality Tests: Hypothetical Data

Switching to the evaluation of the tests using real data in Figure 3.7, the first impression is that significant results are the rule rather than the exception (as in Figure 3.6). This indicates that most of our hypothetical data fulfilled one of their requirements: They represented rather extreme cases one is usually not faced with in reality.

All tests produced significant results on the $\alpha_1 = 0.95$; $\alpha_2 = 0.99$ levels for the Danish Register Data and for Respiratory Diseases. As both data-sets show a pronounced sinusoidal pattern, it is obvious that the nonparametric tests yield this result. The significant values for the Goodness-of-Fit tests and the "Edwards' Family" when evaluating the Danish register data underlines their sample size dependency: If one were taking simply the relative monthly frequencies, the Danish data would show less fluctuation than the hypothetical sine wave which was tested negatively for small sample size. The χ^2-Goodness-of-Fit-Test, especially, seems to be extremely sensitive to sample size. It does not yield significant results for the monks data at all, while the nuns data are highly significant. When looking at the histograms of both data (Figure 3.4 b and c), the eye would assign the tag "seasonal" to the monks' rather than to the nuns' monthly distribution of deaths. For the five data-sets shown in the lower part of Figure 3.4, the nonparametric tests show only rarely significant results. This is probably due to the sparse data of some data-sets such as Lymphoma (Figure 3.4 h) or Leukemia (Figure 3.4 j) where assigning ranks might not be the best option.

Most of the desired properties for inequality indices do not narrow down the choice for a "best" seasonality test. Tests which are based on ranks like the nonparametric tests presented here fulfill the "relativity principle". According to that principle, the outcome should be dependent on the relative contribution of each group — regardless of the sample size. On the contrary, the nonparametric tests cannot pass the "transfer principle". If deaths were "shifted" from months with low mortality to months with high mortality, the non-parametric tests would not necessarily result in more significant ρ-values. This would, however, be the case for the Goodness-of-Fit tests and the "Edwards Family". "Standardization" poses no problem for any of these tests. They are, by definition, designed to return values between 0 and 1 for ρ. All tests are relatively "easy to comprehend, intuitively meaningful and easy to explain to others" [126, p. 11] (Intelligibility): the Goodness-of-Fit tests analyze if an observed distribution deviates too much from a hypothetical distribution which cannot be explained by chance. The tests based on Edwards' contribution have some kind of geometrical framework, where the deviation from a uniform distribution is tested. The nonparametric tests examine whether the observed data show a peak-period of either 6, 5, 4, or 3 months, respectively. The favorable properties of *Sensitivity* and *Robustness* [126] have not been introduced before. If data are described with one statistic, the first choice is often a measurement of the central tendency. Typical examples are the mean and the median. While the mean is often the preferred description, one has to be aware that it is not very robust when the data contain outliers. Likewise, some seasonality tests could be also prone to be too sensitive when faced with some outliers. Nevertheless, seasonality indices should also not be too robust: If there is one extreme outlier, for example caused by an influenza epidemic, a reasonable test should not treat this as similar to another value which might be just slightly higher than values in any other month. Thus, a seasonality index based purely on ranks is too robust. "Sensititivity" and "Robustness"

Real Data (Part A)												
	Wrigley		Nuns		Monks		Union		Danish		Resp.	
Test	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
1. Goodness-of-Fit-Tests												
1.1 Chi-Square - Goodness of Fit	-	-	+	+	-	-	+	+	+	+	+	+
1.2 Kolmogorov-Smirnov-Type-Statistic	-	-	+	+	+	+	+	+	+	+	+	+
2. Edwards' Family												
2.1 Edwards' Test	-	-	+	+	+	+	+	+	+	+	+	+
2.2 Roger's Extension of Edwards' Test	-	-	+	+	+	+	+	+	+	+	+	+
2.3 Pocock's Method	-	-	+	+	+	-	+	+	+	+	+	+
2.4 Cave and Freedman	-	-	-	-	-	-	-	-	+	+	+	+
3. Non-Parametric-Tests												
3.1 Hewitt's Test [1]	+	-	+	+	+	+	+	+	+	+	+	+
3.2 Rogersons' Generalization for												
3.2.1 a 5-months peak [1]	+	+	+	+	+	+	+	+	+	+	+	+
3.2.2 a 4-months peak [1]	+	n.a.	+	n.a.	+	n.a.	+	n.a.	+	n.a.	+	n.a.
3.2.3 a 3-months peak [1]	+	n.a.	+	n.a.	-	n.a.	+	n.a.	+	n.a.	+	n.a.
3.3 David-Newell-Test	-	-	+	+	+	+	+	+	+	+	+	+

Real Data (Part B)										
	Anenc.		Lymph.		Suicides		Leukem.		Crohn's	
Test	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
1. Goodness-of-Fit-Tests										
1.1 Chi-Square - Goodness of Fit	-	-	+	-	+	-	+	-	-	-
1.2 Kolmogorov-Smirnov-Type-Statistic	+	-	+	-	+	+	+	-	-	-
2. Edwards' Family										
2.1 Edwards' Test	+	-	+	-	+	+	+	+	-	-
2.2 Roger's Extension of Edwards' Test	+	-	+	-	+	+	+	+	-	-
2.3 Pocock's Method	+	-	+	-	+	+	+	+	-	-
2.4 Cave and Freedman	-	-	-	-	-	-	-	-	+	+
3. Non-Parametric-Tests										
3.1 Hewitt's Test [1]	-	-	-	-	-	-	-	-	-	-
3.2 Rogersons' Generalization for										
3.2.1 a 5-months peak [1]	-	-	-	-	+	-	-	-	-	-
3.2.2 a 4-months peak [1]	-	n.a.	-	n.a.	-	n.a.	-	n.a.	-	n.a.
3.2.3 a 3-months peak [1]	-	n.a.	-	n.a.	+	n.a.	+	n.a.	-	n.a.
3.3 David-Newell-Test	+	-	+	-	+	-	+	-	-	-

1) The actual levels of significance for the non-parametric tests are:		
Levels of Significance	0.05	0.01
Hewitt	0.0483	0.0130
Rogerson 5 months peak	0.0562	0.0152
Rogerson 4 months peak	0.0470	n.a. (Max. Ranksum=42; p_{42}=0.0267)
Rogerson 3 months peak	0.0545	n.a. (Max. Ranksum=33; p_{33}=0.0545)

Fig. 3.7. Results for Seasonality Tests: Real Data

are excluding principles. The nonparametric tests are very robust against outliers. Consequently, they cannot be too sensitive for sudden, abrupt changes in the distribution. The other two groups of tests behave exactly the other way around.

Our analysis does not yield "the best seasonality test". Depending on data and the relevant research question, different tests are useful. One should always keep in mind that some tests are quite sensitive to sample size. Another important feature is the distribution of the underlying data: Do we have a relatively smooth pattern or do the data look rather erratic? Last but not least, the test should be also aimed at the research question: Do we assume that the underlying data have a bimodal pattern? Only in that case, the test developed by Cave and Freedman [44] can be recommended. If it is expected that the disease/cause of death has a rather sudden prevalence throughout the year for a relatively short period of time, Rogerson's generalization of Hewitt's test for 3, 4 or 5 months should be used. In the case of smooth data structure across the twelve months, it is probably best to use Hewitt's test. As it is based on ranks, it would be probably best to use it in conjunction with a seasonality index such as φ_1 to give an indication of the extent of seasonal fluctuations. Goodness-of-Fit tests and "Edwards's Family" should only be used if the data do not show a smooth pattern.

3.6 Evaluation of Time-Series Methods Using Hypothetical Data

3.6.1 Introduction

Evaluating time-series methods aims at a different angle than the discussion of indices and tests discussed above. A general applicable tool should be able to fit a model to data with characteristics one typically observes for seasonal mortality studies [117]. One major part is the correct estimation of the trend component. It is more common in studies of seasonal mortality to have pure count data available than rates of the variable of interest. Thus, a correct estimation of the trend should be flexible enough to incorporate on the one hand changes in the variable of interest. For example, it can be expected that the overall trend in mortality is decreasing over time. On the other hand, compositional changes can push the trend in the other direction. Due to the increased survival chances, for instance, more and more old people are alive which implies an increase in death counts in absolute terms. It should be also obvious that it is necessary for a seasonal analysis of time-series that the seasonal component is not constant over time.

Not all time-series methods discussed before have been analyzed. The "classical decompostion" has been omitted as it assumes a constant seasonal component over time. Instead of X-11 and X-11-ARIMA the latest version, X-12, has been used since it should yield better estimates than the previous version due to improved outlier detection and automatic estimation of ARIMA-Models.

There are not any software package available that contain all remaining time-series methods. Therefore, we had to rely on several packages to investigate the various approaches. Table 3.2 gives an overview which software has been used for which particular method. Besides R [170, 301], we also used Splus, EViews and BV4 [38].

Table 3.2. Software for Implementation of Time-Series Methods

Method	Software	Version
X12	EViews	4
SABL	Splus	2000
STL	R	1.8.1
TRAMO/SEATS	EViews	4
BV4	BV4	4.1

3.6.2 Description of Data-Sets

In contrast to seasonality indices and tests we analyzed time-series methods only with hypothetical data. Real data are used in Chapter 4.

We used seven synthetically generated data-sets with an increasing level of complexity. The construction of these data is briefly outlined in Table 3.3. Seasonal rates are rarely available. This is why we wanted to reflect this fact in our hypothetical data by constructing them as count data. We started with a simple model being constant in the trend and the seasonal component. No residuals are put into the data (Model I). It should be expected from any seasonal decomposition/adjustment procedure to extract the trend and the seasonal component correctly. For any subsequent model (Models II–VII) we introduced a third-order polynomial to obtain a monotonously increasing trend. Starting with Model IV we modeled a linearly increasing seasonal component. The last models' seasonal components employ also a second, semi-annual wave in the data. This should test whether the seasonal procedures are also able to detect heat-related deaths during summer. We chose three distributions from which the data are drawn: (1) none for Models I, II, and IV; having no residual component at all is very unlikely in reality; (2) therefore models III, V, and VI followed a Poisson distribution; however, the Poisson distribution is sometimes inappropriate. This can be easily seen if the requirement of the Poisson distribution of $\mathrm{E}(x) = \mu(x) = \mathrm{Var}(x)$ is not met. One often encounters so-called overdispersion ($\mathrm{Var}(x) > \mathrm{E}(x)$). This can be typically caused by unobserved heterogeneity. As we use only time as a covariate it can be assumed that this proxy is unable to catch all significant influences and, as a consequence, we are faced with unmeasured factors. A pure Poisson process is therefore the exception rather than the rule. Thus, (3) we opted to use a Negative Binomial distribution [22, 41, 292].[14]

3.6.3 Results and Discussion

There are different approaches to evaluate statistical methods. We decided to base our judgment on visual inspections of the decomposition process. While

[14] While a Poisson distribution requires $\mathrm{E}(x) = \mathrm{Var}(x) = \mu$, the negative binomial distribution relaxes the assumption about the variance with $\mathrm{E}(x) = \mu$ and $\mathrm{Var}(x) = \mu + \frac{\mu^2}{\theta}$ [389]. In our application, we set θ to 100.

Table 3.3. Hypothetical Time-Series Data

Model	Trend	Seasonal Component	Errors
I.	constant	constant	—
II.	monotonously increasing	—"—	—
III.	—"—	—"—	Poisson
IV.	—"—	increasing	—
V.	—"—	—"—	Poisson
VI.	—"—	increasing "heat-related mortality"	—"—
VII.	—"—	—"—	Neg. Binom.

a theoretical statistician may criticize this, the major advantage is that one can immediately recognize whether a specific method caught the important characteristics of the underlying data. The following Figures 3.8–3.14 show the results for the Models I–VII described before. For all our calculations we did not use the default settings but tried to adapt the methods as closely as possible to the actual data. In the case of X-12, for example, we linked the components multiplicatively or log-additively according to our initial assembling of the data. In real world applications, one does not have that background knowledge. Therefore, the results for X-12 might show better results for our hypothetical data than for real world data. TRAMO/SEATS did not pose any problems for the implementation, nor did SABL or the Berliner Verfahren. Applying STL was less straightforward: As pointed out in the original paper [48], there are 6 parameters to be entered into the model. Five of them can be found automatically (e.g. number of observations), for one parameter, however, there is no straightforward solution. Unfortunately, it is a crucial parameter for our purposes: the smoothing parameter for the seasonal component. We followed Cleveland et al.'s suggestion to visually inspect various parameter values [48]. Our analysis resulted in an optimal value of approximately 7 for all our models. Lower values made the seasonal component change too quickly, higher values resulted in seasonality being too smooth. [15]

The column on the left in each figure represents the "real" data (i.e. the input). Combining the trend (f) with the seasonal component (k) and the residuals (p) resulted in the "real data" (a). Those "real data" were used as input for the four different seasonal decomposition methods X-12, SABL, STL, and TRAMO/SEATS and the Berliner Verfahren ("BV4"). Perfectly working methods should decompose the input data in exactly the components we used for the composition initially. We can see the outcome of these methods in columns 2–6 in each graph for X-12 (column 2), SABL (column 3), STL (column 4), TRAMO/SEATS (column 5) and BV4 (column 6).

[15] Cleveland et al. advise to use odd numbers ≥ 7 [48]. We actually searched values from 1 until 50 wheres the original authors looked only from 7 until 35.

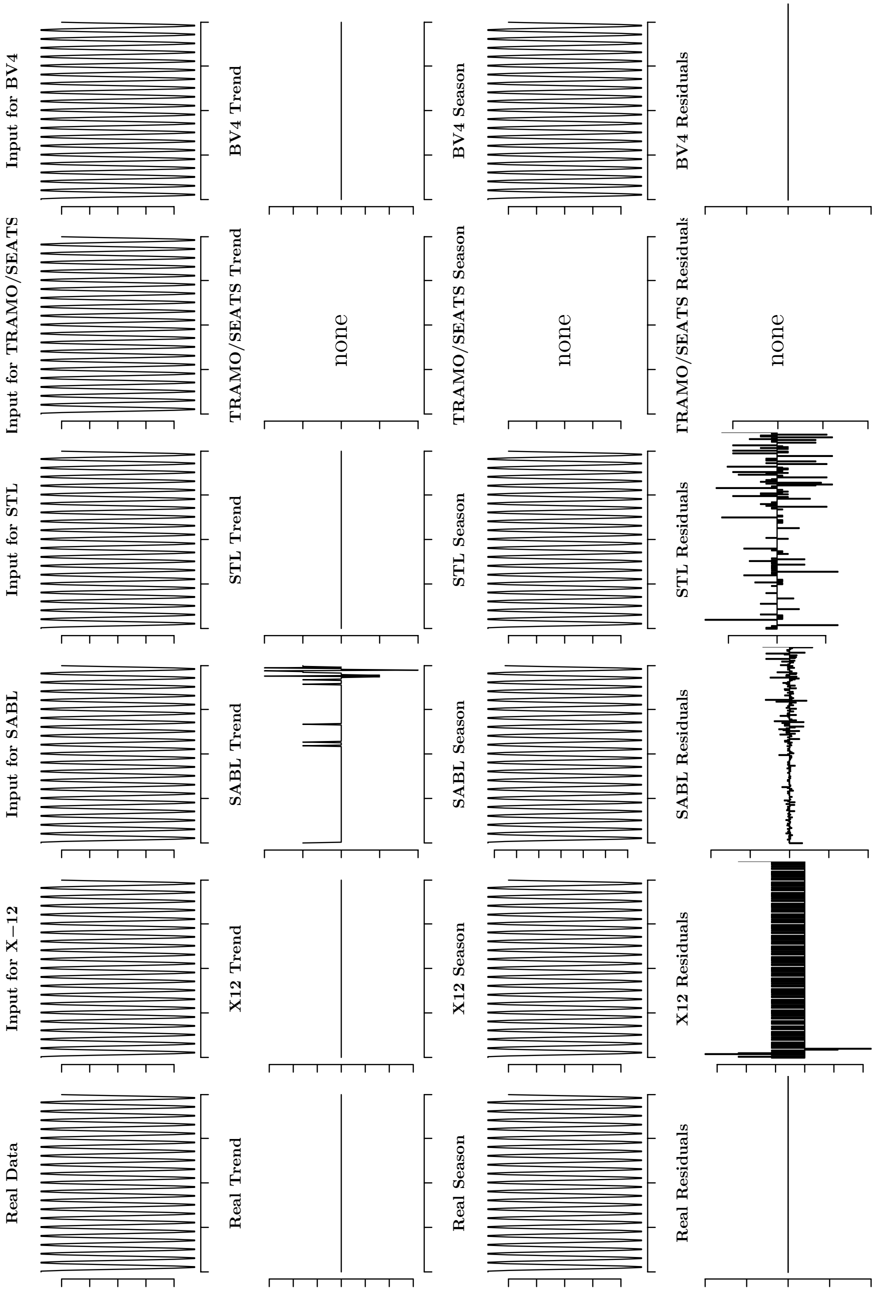

Fig. 3.8. Seasonal Decomposition of Time-Series — Model I

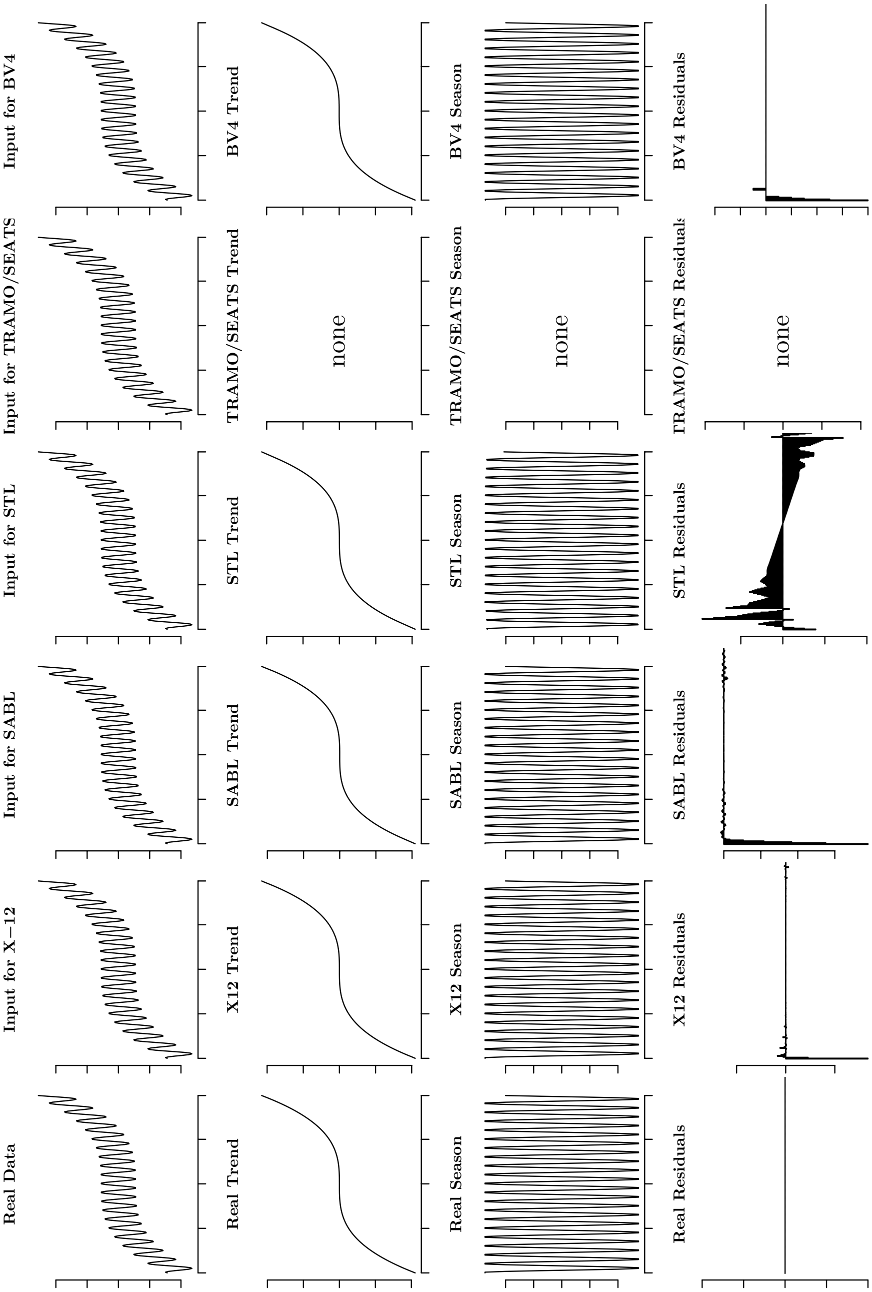

Fig. 3.9. Seasonal Decomposition of Time-Series — Model II

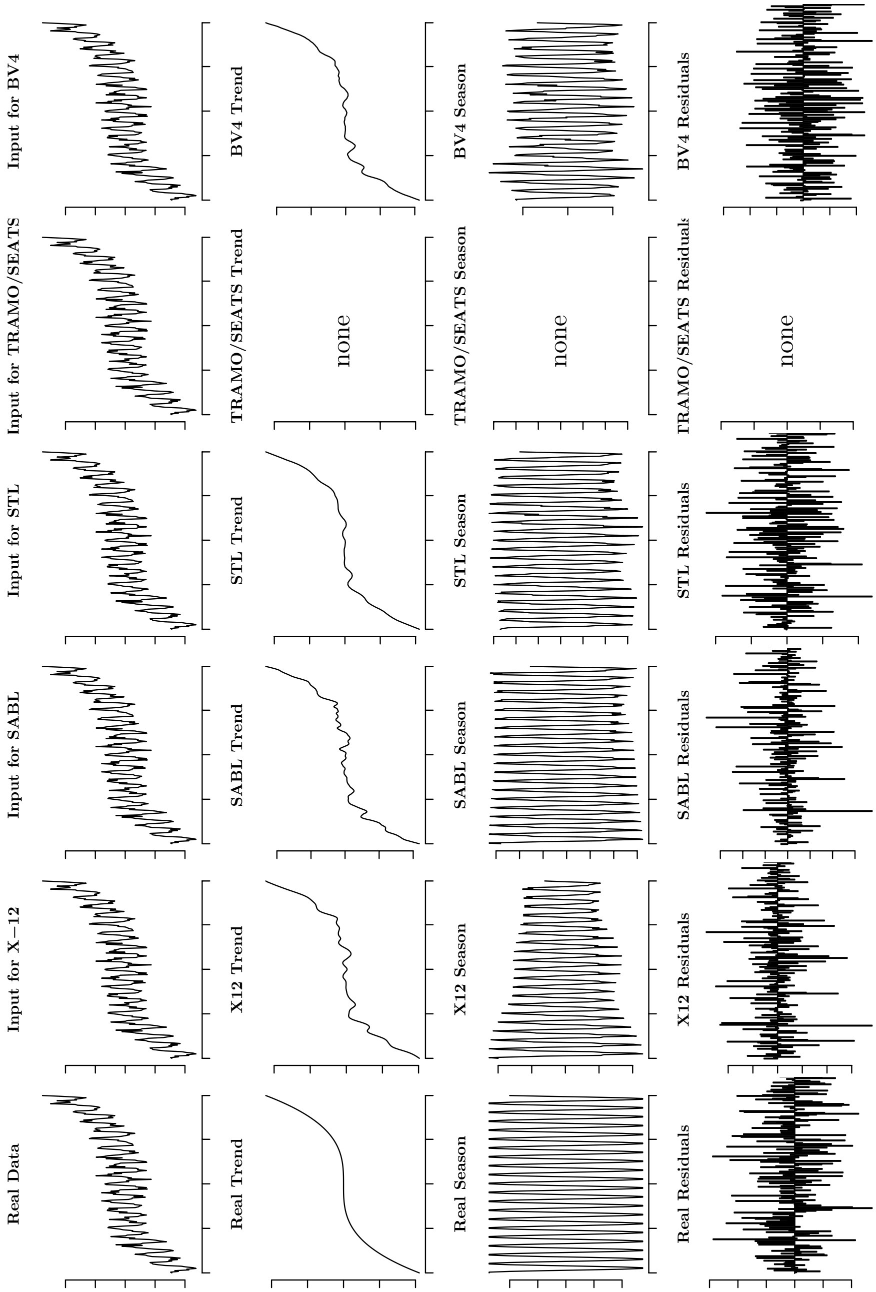

Fig. 3.10. Seasonal Decomposition of Time-Series — Model III

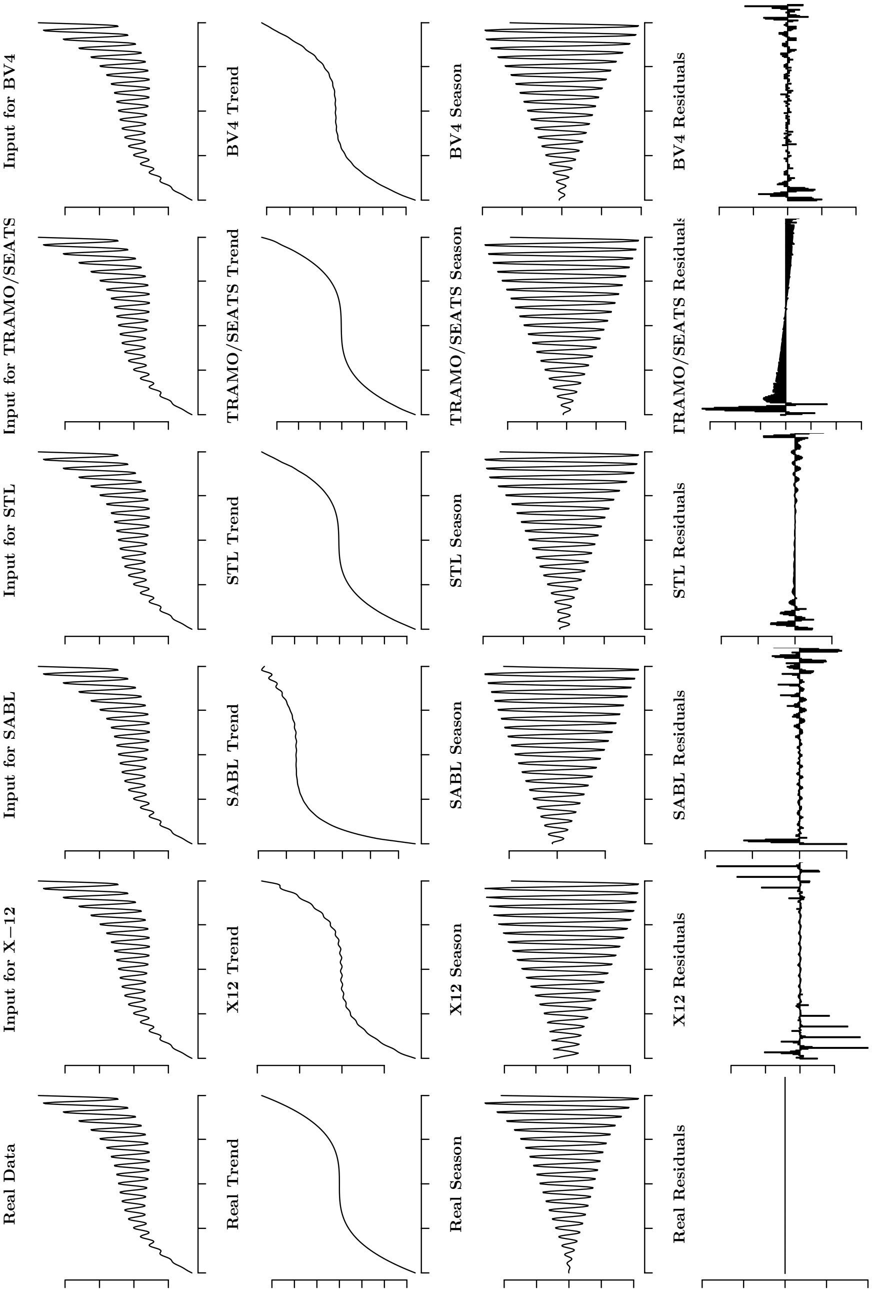

Fig. 3.11. Seasonal Decomposition of Time-Series — Model IV

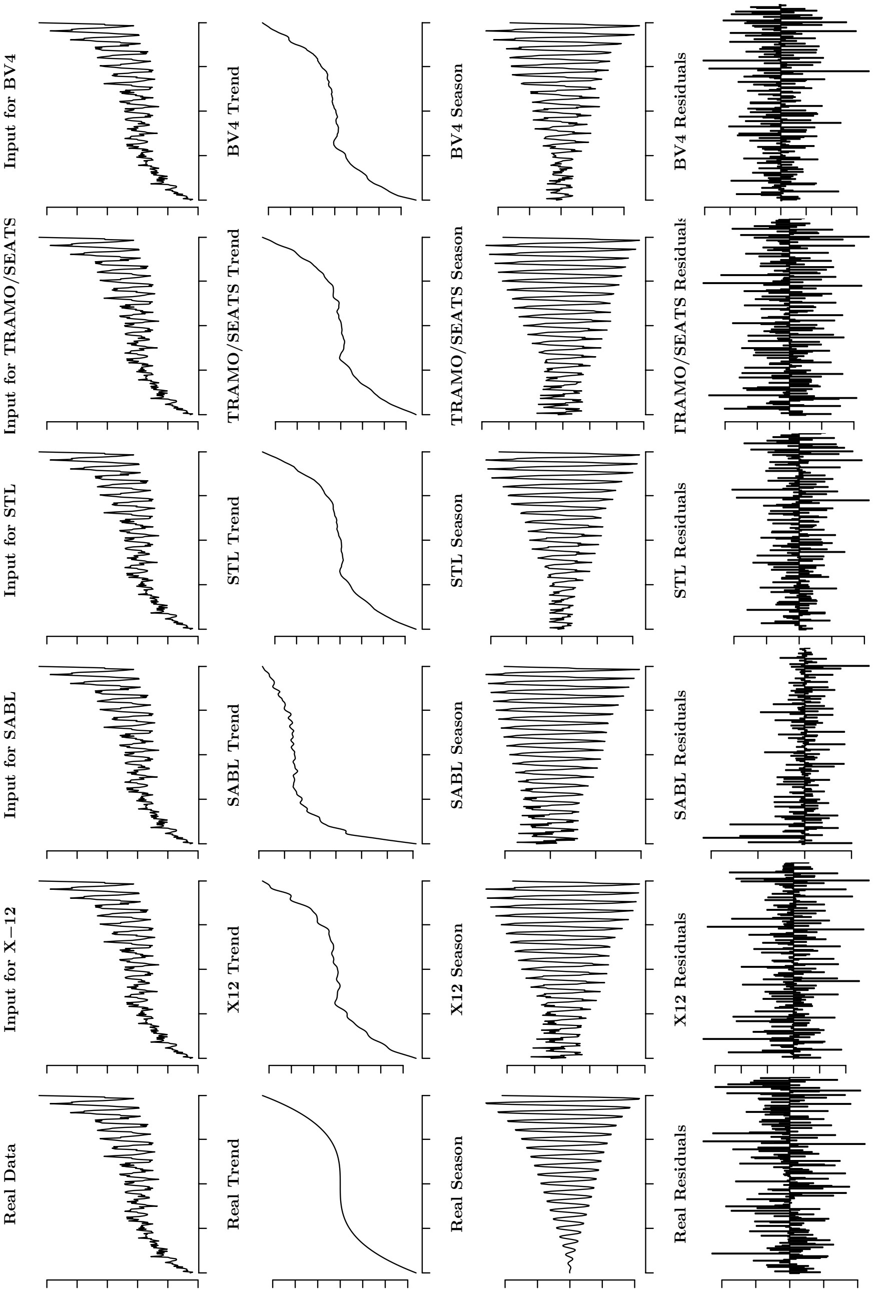

Fig. 3.12. Seasonal Decomposition of Time-Series — Model V

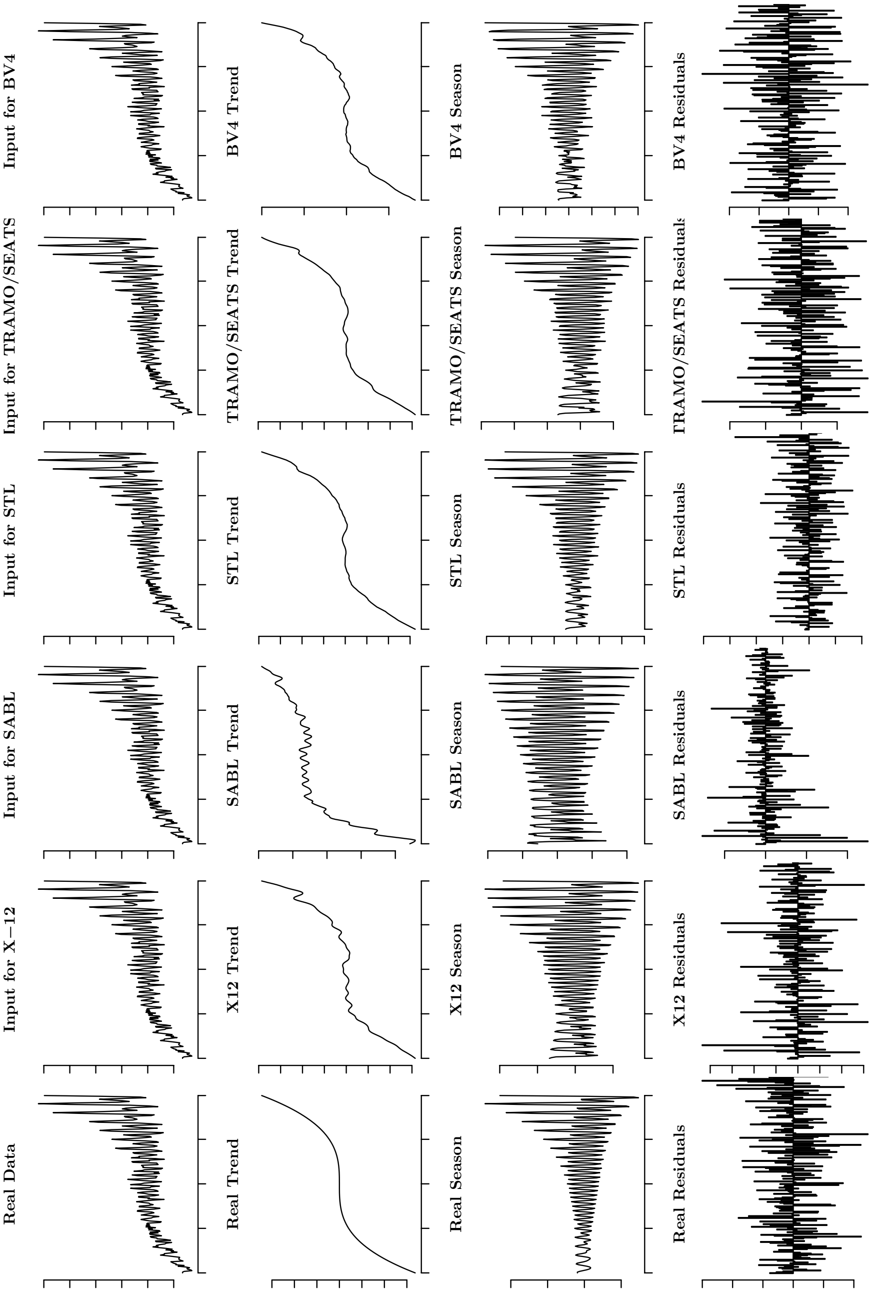

Fig. 3.13. Seasonal Decomposition of Time-Series — Model VI

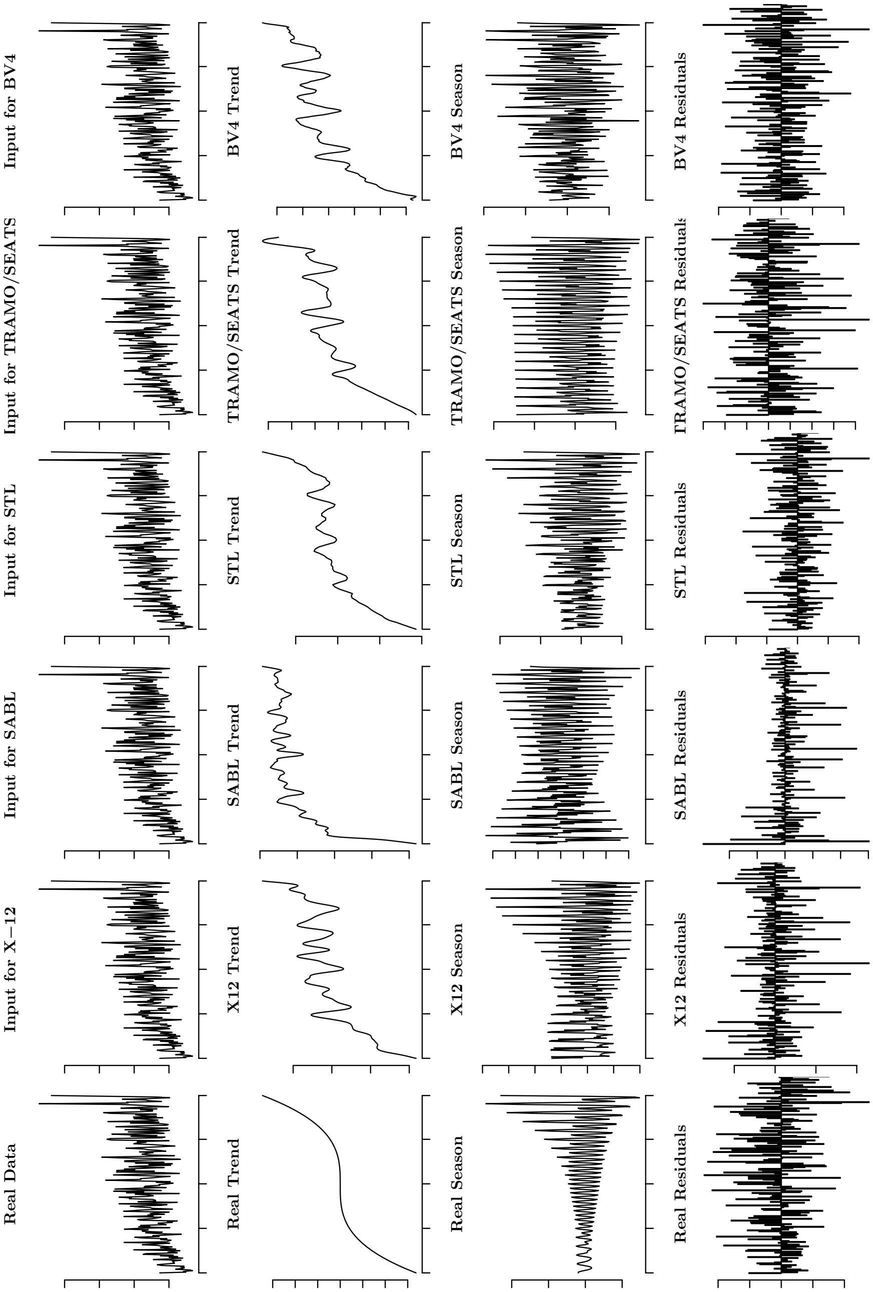

Fig. 3.14. Seasonal Decomposition of Time-Series — Model VII

For Model I (Figure 3.8), BV4 works perfectly; the methods X-12, SABL and STL perform almost as well as the procedure from the German Statistical Office. This close fit should be expected anyway, as a constant trend and a constant seasonal pattern represents the easiest seasonal pattern. The implementation of TRAMO/SEATS in EViews 4.1 did not work for Models I–III. Surprisingly the failure of this method is highly correlated with a constant seasonal pattern. Due to the lack of detailed information [299], it was impossible to determine the reason for the program's crashes.

With the exception of TRAMO/SEATS, the four other methods performed again very well for Model II. Solely STL's extraction of the residuals was slightly problematic: While no residuals should appear, STL, nevertheless extracted residuals. In addition, those residuals are highly auto-correlated, indicating that important characteristics of the data are misspecified into the irregular component. However, this is only a minor drawback: for reason of simplicity, no numbers have been put on the scales. The mean value of the trend is 400 and the amplitude of the seasonal component is 54. The residuals' mean amplitude height is 0.40 and their maximum value amounts only to 1.92. This misspecification of the irregular component is, thus, rather negligible.

Seasonal decomposition by standard methods becomes tricky when artificial noise is added to the data, as shown in Model III when the data are drawn from a Poisson distribution (Figure 3.10). All procedures contain a somehow wiggly trend. Only SABL seems to extract the seasonal component very well. X-12 shows a decline in seasonality; STL's algorithm produces a fairly shaky result. The seasonal component of the Berliner Verfahren is the least stable.

Model IV (Figure 3.11) is the first model for which TRAMO/SEATS was working. STL, TRAMO/SEATS and BV4 reproduced the trend almost identically to the input data. X-12's of the underlying third-order polynomial is only slightly worse, whereas SABL's trend is smooth but estimated wrong.[16] All methods performed remarkably well for the extraction of the seasonal signal.

The quality of the four decomposition methods declines rapidly, starting with Model V (Figure 3.12). Besides a monotonously increasing trend and seasonal component we allowed the data again (Model III) to be derived from a Poisson distribution. SABL still faces the same problems when plotting the trend as in Model IV. But the other methods (X-12, STL, TRAMO/SEATS, BV4) also do not show a clear signal extraction for the trend; it becomes rather wiggly. None of the decomposition methods is able to mirror the seasonal component exactly into the data. Although all four methods show somehow an increase in seasonality, only STL and BV4 fit the seasonality part relatively well. The results from TRAMO/SEATS, SABL and X-12 are not satisfactory.

So far we have only used one sine and one cosine term to model annual fluctuations in mortality. Model VI (Figure 3.13) introduces a more elaborated seasonal component with a sine and a cosine component of frequency of six

[16] The wrong estimation is not caused by using a log-transform initially and forgetting about re-transforming in the end.

months. This allows to incorporate heat-related mortality (summer excess deaths) into our models. As Model VI is equivalent to Model V with this exception, it should be no surprise that none of the four methods performs better than previously.

Model VII (Figure 3.14) is the most complicated pattern we faced our data with. In addition to a monotonously increasing trend, an annual and a semi-annual ("heat-related mortality") seasonal swing, we input unobserved heterogeneity by drawing our data from a Negative Binomial Distribution with a relatively low value of the dispersion parameter Θ.[17] None of the five methods is able to capture the trend or the seasonal component even remotely. All trend estimates show a wiggly upward tendency but neither X-12, SABL, STL, TRAMO/SEATS, nor BV4 mirror the underlying third-order polynomial correctly. Furthermore, the seasonal component is not extracted properly by any of the standard methods: X-12, SABL, TRAMO/SEATS and BV4 seem to be inadequate. The general approach of STL seems to work well for seasonality. Its estimate of this component is, nevertheless, too shaky to be declared satisfactory.

Thus, evaluating time-series methods with hypothetical data did not result in one procedure which can unanimously be recommended. For simple data patterns, the standard methods yield satisfactory results. If these approaches are, however, faced with data structures one can typically encounter in demography (i.e. variable trend, changing seasonality, overdispersion), none of them extracts the entered components well enough. We rather suggest, therefore, the method outlined in Chapter 4 which is especially tailored for those situations and returns the trend as well as the seasonal component almost identical to the simulation input.

3.7 Summary

The aim of this chapter was to present and critically evaluate indices, tests and time-series methods for seasonality. For that purpose various methods which are used in the literature have been presented, discussed and evaluated with hypothetical (indices, tests, time-series methods) and with empirical (indices, tests) data.

Three indices were presented: a winter/summer ratio, a dissimilarity index and a measurement based on entropy. Among them, the winter/summer ratio seems to be the best choice, mainly because of its easy interpretability and that it takes the ordering of the months into account.

Recommending a test for seasonality is less straightforward. Several tests have been presented and discussed which can be categorized in three classes: Goodness-of-Fit tests, the "Edwards' family", and nonparametric tests. Choosing an appropriate test should be guided by the underlying research question

[17] The lower the dispersion parameter Θ, the larger the variance of the data: $\mathrm{Var}(Y) = \mu + \frac{\mu^2}{\Theta}$ [cf. 389].

and by the nature of the data. For the “normal” application, i.e. a smooth pattern with one peak during the year, Hewitt’s test is probably best [150]. Because this test is purely based on ranks, it should be used in conjunction with the winter/summer-ratio to have a measurement also of the height of the seasonal fluctuations. Generalizations of Hewitt’s test [315] can be employed if one assumes sudden outbreaks of certain diseases throughout the year which last only a limited amount of time. If two peaks during the year are expected such as for Crohn’s disease, the test proposed by Cave and Freedman seems to be appropriate [44]. If the data are rather erratic, one should use either one of the Goodness-of-Fit tests or one from the “Edwards’ family” [e.g. 84].

Five common time-series methods (X-12, SABL, STL, TRAMO/SEATS, BV4) have been evaluated using seven models of simulated data with increasing complexity. The general outcome is not convincing: If any of those methods are faced with complicated data, the decomposition of the trend and the seasonal component does not return the input data. For relatively simple simulated data, the signal extraction in all methods works well. The trend and the season in the given data, and after the decomposition process, are almost identical. Sudden changes in the trend does not pose any problems. Problems arise on the one hand if the seasonal pattern is not constant over time. Methods which are unable to handle this, can not be applied as changes in the seasonal component over time (or age) is often the main interest in seasonality studies. On the other hand, the evaluated time-series methods fail to return the entered signals if the data are derived from a Poisson distribution or from a Negative Binomial distribution. In practice, especially the latter distribution appears to be the rule rather than an exception if data are not rates but counts and if relevant factors are unmeasured. It is difficult to point at the exact estimation problem of these standard methods as they are quite complicated due to the filters employed and the various iterative steps involved.

Due to these shortcomings, a new method has been developed which is able to incorporate changes in the trend, the seasonal component and unobserved heterogeneity. This novel approach is presented, evaluated and applied to real data in Chapter 4 (page 83).