Point processes in traffic safety analysis

R-94-51
Frits Bijleveld
Leidschendam, 1994
SWOV Institute for Road Safety Research, The Netherlands

SWOV Institute for Road Safety Research
P.O. Box 170
2260 AD Leidschendam
The Netherlands
Telephone: +31 70-3209323
Telefax: +31 70-3201261

Summary

Presently, the usual method of analyzing accident data in time is through the analysis of the sequence of accident counts. Usually, the number of accidents per month or even per annum is used. Results of this kind of analysis are influenced by the starting point of such a sequence and by the length of the intervals used. The aim of this study is to investigate the possibilities for analyzing accident data independently of the choice of the length and consequently independently of the choice of the starting point. This seems to be possible using the original points of time as recorded.

Techniques developed are based on the *Doob-Meyer* decomposition of the stochastic process of the count of accidents. It is found that many techniques are readily available.

It is assumed that the accident process has an intensity process. Under certain regularity conditions it is found that such an intensity function exists. It is attempted to build a model based on an exponential variant of a Fourier system that estimates that intensity function.

Some extensions, covering exogenous variables and intervention analysis are discussed. Finally, some simulations and a real life problem are given.

It is found that the current implementation suffers from a non-optimal goodness-of-fit criterion and lacks the ability of inclusion of exogenous variables. Apart from this, the Fourier system may be extended, possibly by wavelets.

Contents

Preface

The main purpose of this work is to write a thesis in statistics, needed to complete my studies in mathematics at Leiden University, Netherlands. The second reason to write *this* thesis was some dissatisfaction with the techniques available to me at my (already) long standing work at SWOV, a road safety research institute in the Netherlands. The idea was triggered at a discussion on the effects of seatbelts in cars. This discussion resulted in figure 1. As far as the discussion is concerned, the results of this thesis have not been applied in that direction yet. Most countries introduced seatbelt legislation long before the accident times were recorded reliably enough. Other countries, like Italy, have introduced legislation in recent years together with other measures.

Another aim was to get more sensitive methods. It is not yet clear if this aim has been reached, this should become apparent from empirical evidence.

I wish to thank my tutor Dr Sara van de Geer very much for her help, patience and endurance, I could not have done without. Apart from her many thanks are directed to the library service of SWOV, in particular Dennis van der Braak, for I cannot do without them as well (not only for this thesis). I also want to thank my colleagues at SWOV (in particular Dr Peter Polak) for their suggestions and SWOV for letting me use accident data.

Leiden, 1993.

1. Introduction

1.1 The problem

In the Netherlands an extensive procedure is conducted to register accidents occurring on roads. Among many characteristics the point of time at which the accident occurred is recorded. This point of time is of course an approximation. Theoretically, the accuracy is up to one minute. In practice, however, most accidents are recorded at a precision of about fifteen minutes.

The question arises whether it is possible to use this information in order to analyze the development of the occurrence rate of some kind of accident. Currently, at best data consisting monthly counts are analyzed. This kind of method using the number of accidents for a sequence of periods of time of a particular length is vulnerable to two kinds of problems, although sometimes no alternatives exist. Both problems are similar to the problems using histograms in density estimation (see Härdle [1990]). Both the choice of length and the starting point influence the results of an analysis. Sometimes these effects can be decisive (see figure 1).

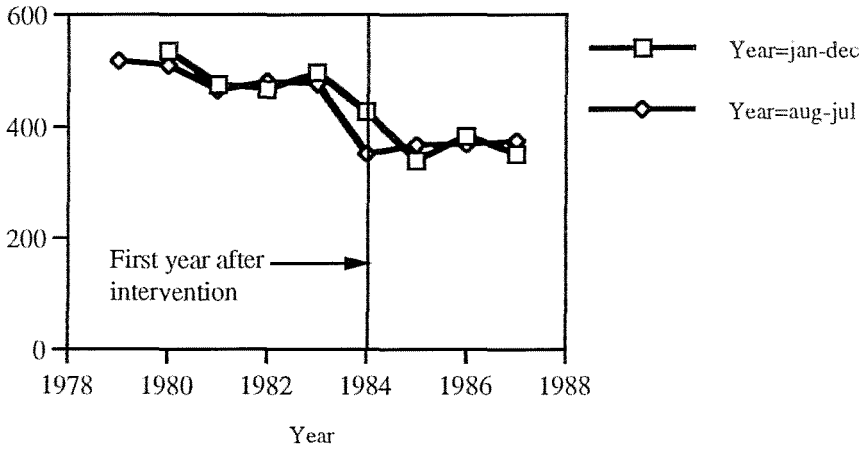

Figure 1: *Two alternative registrations of killed car occupants in West Germany.*
In August 1984 the German government introduced fines for not using safety-belts. This resulted in a sudden large increase in belt-usage. It was expected that this increase would result in a similar sudden decrease in the number of killed (front seated) car occupants. One series is the usual annual number of killed car occupants, the other is a reordered series consisting of the number of killed car occupants aggregated over the months August through July every year. It is clearly visible that the traditional annual counts obscure a possible effect. Although the process of killed car occupants is more complex than the process of the accidents themselves, the example should be informative. (Source:Statistisches Bundesamt [1988]).

In addition, a problem might be the obscuring of effects within the sample period. In particular, this can be important when an (intervention) analysis is conducted on an intervention that influences only a part of the day, specially when

it is not (yet) very well known how. More generally, a change in the structure of a periodic can be of interest by itself.

The aim of the present study is to investigate the possibilities for analyzing accident data independently of the choice of the size and consequently independently of the starting point. This seems to be possible using the original points of time as recorded.

1.2 Accident data

On the theoretical side, one is tempted to assume that at any point in time (or better: any time interval) a positive probability exists that an accident of some kind occurs. One should realize that this probability of the occurrence of an accident of a particular type is not constant over time. The possibility of an accident is heavily dependent on many variables, such as the presence of traffic in that interval, visibility conditions and so forth. Because these conditions may vary easily, it seems necessary to assume that the accident probabilities also vary over time. Secondly, because conditions may change seemingly at random, it is attractive to assume the probabilities vary randomly, or at least to some extent. Furthermore, it is noted that while probabilities of 'exotic' accidents are slim, they are still nonzero! These assumptions play a central role in the following discussion.

Additional assumptions are:
– The casualties as a result of the accidents do not significantly affect the population.
– No more than one accident can happen at (exactly) the same time.
– The registration system is capable of observing accidents at any (realistic) rate. Registration will not be overrun causing periods where accidents can by pass the registration procedure without being registered because the system cannot handle them. In practice however, this assumption is violated to some extent, mainly due to police workload.
– There is not a so-called after effect: when an accident occurred, no period follows in which no other accident can occur.
– Accidents do not influence each other. When an accident causes another accident both are seen as one accident. This is implemented in practice in the Netherlands, although in the case of a 'pile-up', cars colliding later are usually seen as new accidents.

It is assumed that the accident process, the process that seems to 'generate' accidents, is a result of:
– The traffic volume.
– The traffic participants.
– Coincidental circumstances such as:
  - road works.
  - weather influences.

The traffic volume will result in a sort-of *base(line) intensity process* of the number of traffic participants at risk at a certain time. In theory this process could be observable, but even in a small country such as the Netherlands this is practically impossible. This process, together with some (generally unknown) danger process is supposed to generate a process that 'generates' dangerous situations, that may or may not result in accidents of some kind.

In this work, aspects of this combination process, the accident process, are studied. It is thought that this process has a sort of underlying intensity process that is assumed having continuous paths, or at least it is stochastically not distinguishable from a continuous path process. This intensity process, further called intensity function, is the main object of study in this work.

Another, more serious note is that for some types of accidents, the reliability of the registration is lower than for other types, mainly due to differences in damage (personal and material) and insurance matters (no-claim rebates). This results in the fact that the accident process is only partially observed. This matter is addressed too.

2. Theoretical background of the intensity function

2.1 Definitions

Before starting, a few notations and definitions should be agreed upon. First, it should be pointed out that only processes in one dimension are being considered. Furthermore, the point-processes will be considered to run between 0 and 1, or other simple margins when appropriate. Because of the simplicity of transforming between those various bounds the text will switch between those bounds for convenience.

Definition 2.1. The point process $N_t, t > 0$ is defined as the number of accidents (or events) that have occurred up to the point of time t, $N_t = \sum_i I(\tau_i \leq t)$. Naturally, $N_t \in \mathbb{N} \cup \{0\}$. A *marked* point process is a process in which every point has a mark, that is a property. In the traffic safety context this mark may indicate the *kind* of accident that occurred at time τ_i. Finally a *p-thinned*-point process is a process $N'_t = \sum_i U_i I(\tau_i \leq t)$, U_i conditionally independent given N. $P(U_i = 1|N) = 1 - P(U_i = 0|N) = p$. p can be seen as the retention probability of a point τ_i. In the traffic safety context this is the probability of an accident actually being recorded. Obviously, the intensity of the thinned point process is p times the intensity of the parent process.

2.2 The Doob-Meyer theorem

The Doob-Meyer Decomposition [Karr, 1991, section 2.3] offers the theory of the existence of a so called compensator for every point process N_t. This compensator serves as the integral of the intensity function we are looking for. This theorem cumulates in [Karr, 1991, Theorem 2.22] that states that for a point process satisfying the assumptions:

Assumption 2.1. Let the point process N_t induce the filtration $\mathcal{F}_t$ in the probability space $(\Omega, \mathcal{F}, \mathcal{P})$. Assume the filtration $(\mathcal{H}) = \mathcal{H}_t$ has the following properties:
- $\mathcal{H}$ is right continuous.
- $\mathcal{H}_0$ contains all $\mathcal{P}$-null sets in $\mathcal{H}_\infty = \bigvee_{t \geq 0} \mathcal{H}_t$.
- N is adapted to $(\mathcal{H})$, so $\mathcal{F}_t \subseteq \mathcal{H}_t$
- $E(N_t) < \infty$ for each relevant t.

Theorem 2.1. *There exists a unique random measure Φ on $\mathbb{R}^+$ such that:*
- *The process Φ_t is predictable with respect to $(\mathcal{H})$.*
- *For each nonnegative predictable process C:*

$$E[\int C dN] = E[\int C d\Phi] \tag{2.1}$$

Φ is called the compensator of N. Not mentioned in [Karr, 1991, Theorem 2.22] but mentioned in [Fleming & Harrington, 1991, Theorem 1.4.1] is that Φ_t is an increasing process.

Corollary 2.2. *Using* $C \equiv 1$*:*

$$E[\int_0^t dN_s] = E[\int_0^t d\Phi_s]$$

Theorem 2.3. *[Karr, 1991, Theorem 2.14] the following statements are equivalent:*
– Φ is a compensator of N.
– The process $M_t = N_t - \Phi_t$ is a (mean zero) martingale, with respect to the filtration $(\mathcal{H})$.

In the case of thinning, the compensator of the p-thinned process is $p\Phi_t$.

It can be noted, as many authors do, that the point process N_t can thus be decomposed in a predictable process Φ_t and a non-predictable process M_t. What can be predicted, in Φ_t, is mainly determined by the amount of information in $(\mathcal{H})$. Karr states [Karr, 1991, section 2.4] that the existence of a *stochastic* intensity depends on the amount of information in the filtration. In the extreme case that the filtration contains all possible information, the compensator will be non-stochastic and equivalent to the point process, thus non-continuous. The point process will be non-stochastic too in this extreme case.

The next step might be to study under what circumstances (Φ) has a (stochastic) intensity ϕ with respect to a (Lebesque) measure, defined by:

$$\Phi_t = \int_0^t \phi_s ds \tag{2.2}$$

The mere existence of Φ through theorem 2.1 is not sufficient for the existence of an intensity function. What is left to prove is that it is reasonable that (Φ) has a version that is almost surely differentiable. Sufficiency follows from the assumption that N is simple, i.e. no two accidents can happen at the same time. This assumption was already stated in the introduction. Heuristically, as a consequence of equation 2.1 using [Karr, 1991, proof of theorem 2.14]:

$$C(u, \omega) = I_{(s<u\leq t)} I_{(\omega \in \Lambda)}$$

is a non-negative, predictable process, so $E(N_t - N_s|\Lambda) = E(\Phi_t - \Phi_s|\Lambda)$. This can be rewritten to $E(dN_t|\Lambda) = E(d\Phi_t|\Lambda)$ for all $\Lambda \in \mathcal{H}_{t-}$. Because (Φ) is predictable, thus Φ_t is measurable with respect to $\mathcal{H}_{t-}$, it follows:

$$d\Phi_t = E(dN_t|\mathcal{H}_{t-}) \tag{2.3}$$

Thus ϕ_t is defined though:

$$\begin{aligned} \phi_t dt =& E\left(dN_t|\mathcal{H}_{t-}\right) \\ =& P\left(dN_t = 1|\mathcal{H}_{t-}\right) \\ =& P\left(dN_t > 0|\mathcal{H}_{t-}\right) \end{aligned} \tag{2.4}$$

Definition 2.2. When existent, the intensity process is ϕ_t is defined through $\Phi_t = \int_0^t \phi_s ds$. Due to the fact the Φ_t is increasing, it will be argued that the intensity is nonnegative. In addition the process is supposed to be bounded.

2.3 Consequences of the Doob-Meyer theorem

As regards assumption 2.1, often called the usual conditions, the first two points are generally met. In practice $(\mathcal{H})$ will consist of the accident records and some other, supposedly relevant, information. This results in the fact that the third assumption will be met as wll. The last point is only of theoretical importance. It will be met anytime someone survives to analyze the accident data, but it imposes some constraints on possible models that can be estimated. [Fleming & Harrington, 1991, Theorem 1.5.1] add to theorem 2.1 and theorem 2.3 the result that the process

$$L_t = \int_0^t C_s dM_s \tag{2.5}$$

is a $(\mathcal{H})$ martingale under the conditions above with C a bounded, $(\mathcal{H})$ predictable process. Because $E[L_t] = 0$, this gives:

$$E[\int_0^t C_s dN_s] = E[\int_0^t C_s d\Phi_s] \tag{2.6}$$

which will play a central role together with:

Theorem 2.4. *[Fleming & Harrington, 1991, adaptation of theorem 2.5.4] Under assumptions above (equation 2.5):*
- *The process L_t is a $(\mathcal{H})$-martingale.*
- $EL_t = 0,\ 0 \leq t < \infty.$
- $var(L_t) = E\int_0^t C_s^2 d\Phi_s,\ 0 \leq t < \infty.$

Which can be generalized for $L_{it} = \int_0^t C_{is} dM_s, i = 1, 2$:
- $cov(L_{1t} L_{2t}) = E\int_0^t C_{1s} C_{2s} d\Phi_s,\ 0 \leq t < \infty.$

This theorem may help when tests are to be derived for various martingale estimators.

3. Estimating an intensity function

When a function has to be estimated, an estimation criterion must be decided upon. Often estimators in the sense of *least-squares* or minimum risk are used. If there is no external reason to choose a particular criterion, other considerations are used. These considerations include analytical tractability, numerical efficiency and robustness to some kind of disturbance. Even then quite a number of criteria are available, among them minimizing least-squares, maximizing likelihood, partial likelihood, quasi-likelihood, M-estimators and probably many more. The next (first) subsection is devoted to the maximum-likelihood criterion. This method is employed here. The second section exposes three possible least-squares-like criteria, the latter of which is shown to be equivalent to the maximum likelihood criterion under certain circumstances.

3.1 Derivation of the loglikelihood-function

In this subsection the loglikelihood is derived of both the number of points and their locations. It is shown that this loglikelihood holds for a large class of point processes, not just Poisson processes, because their properties are not used in its derivation.

First, conditionally on n, $P(t_1 = \tau_1, \ldots, t_n = \tau_n)$ is derived:

$$P(t_1 = \tau_1, \ldots, t_n = \tau_n) = \prod_{i=1}^{n} P(t_i = \tau_i | t_1 = \tau_1, \ldots, t_{i-1} = \tau_{i-1}) \tag{3.1}$$

Given ϕ, the likelihood of a realization of the point process up to time t is now computed as follows:
1) The last occurrence $\tau_k < t$ is found. Note that this time is not equal to t.
2) Starting at $\tau_0 \equiv 0$, the likelihoods of the occurrences of τ_{i+1} after τ_i are computed, of course assuming $\tau_1, \ldots, \tau_n$ is an ordered sample.
3) If t is an occurrence time itself, the same as 2) is done for t, otherwise the likelihood of no occurrence between time τ_k and t is computed.
What is needed is a formulation of this likelihood. Define T_k, or for brevity T, as the stochastic variable of the time of the next occurrence. It is assumed that its distribution is continuous, so there is no distinct timepoint that has a positive probability of occurrence. This means that the distribution F can be differentiated. The density f of the distribution is derived next.

It is important to note that the likelihoods are always computed for points (times) in intervals that are open on the left side: $I_k = (\tau_k, \infty)$. Using the assumption that no two occurrences can happen at the same time, this means that for every t there is an open environment that contains no other occurrences. This allows the use of equation 2.3 and through this equation 2.4.

The probability of an occurrence between the times t and $t + \Delta t$ given no occurrences up to time t is:

$$P(t \leq T < t + \Delta t | T \geq t) = \frac{P(t \leq T < t + \Delta t \cap T \geq t)}{P(T \geq t)} = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)}$$

Note that $\{T \geq t\} = \{T < t\}^c \in \mathcal{H}_{t-}$. This allows for the use of equation 2.4 in equation 3.2, further using continuity of F in equation 3.3 it follows:

$$\phi(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{3.2}$$

$$= \lim_{\Delta t \downarrow 0} \frac{P(t \leq T < t + \Delta t)/\Delta t}{P(T \geq t)}$$

$$= \frac{f(t)}{1 - F(t)} \tag{3.3}$$

The right-hand part of equation 3.3 is commonly known as the *hazard*-function. Many sources, including Kalbfleisch & Prentice [1980] give properties of this relation, among them, adapted to the above sketched situation:

$$f(T) = \exp\left(\log\left(\phi(T)\right) - \int_{\tau_k}^{T} \phi(s)ds\right)$$

$$1 - F(T) = \exp\left(-\int_{\tau_k}^{T} \phi(s)ds\right)$$

Putting it all together:

$$L(\phi, \tau_1, \ldots, \tau_n)$$

$$= \exp\left(-\int_{\tau_n}^{\tau_{n+1}} \phi(s)ds\right) \prod_{k=1}^{n} \exp\left(\log\left(\phi(\tau_k)\right) - \int_{\tau_{k-1}}^{\tau_k} \phi(s)ds\right)$$

using $0 = \tau_0$ and $\tau_{n+1} = 1$ or the final time point. This results in:

$$\mathcal{L}(\phi, \tau_1, \ldots, \tau_n) = \sum_{k=1}^{n} \log(\phi(\tau_k)) - \int_0^1 \phi(t)dt \tag{3.4}$$

as the loglikelihood. The maximum likelihood estimator of $\hat{\phi}$ of ϕ is

$$\hat{\phi} = \underset{\phi \in \mathcal{S}}{\operatorname{argmax}}\, \mathcal{L}(\phi, \tau_1, \ldots, \tau_n)$$

with $\mathcal{S}$ a suitably chosen class of functions or sieve, see § 3.3.1 below. In the case of exogenous variables, the marginal likelihood is maximized. Let $\mathcal{H}_t$ be induced by N_t and the exogenous variables Y_t, Y_t predictable. Define for m, $t_i = \frac{i}{m}$, $\Delta N(t_i) = N(t_i) - N(t_{i-1})$ $\Delta Y(t_i) = Y(t_i) - Y(t_{i-1})$ $\Delta\Phi(t_i) = \Phi(t_i) - \Phi(t_{i-1})$ then:

$$P(\Delta N(t_1) = x_1, \Delta Y(t_1) = y_1, \ldots, \Delta N(t_m) = x_m, \Delta Y(t_m) = y_m)$$

$$= \prod_{i=1}^{m} P(\Delta N(t_i) = x_i | \Delta N(t_1) = x_1,$$

$$\Delta Y(t_1) = y_1, \ldots, \Delta N(t_{i-1}) = x_{i-1}, \Delta Y(t_{i-1}) = y_{i-1})$$

$$= \prod_{i=1}^{m} P(\Delta Y(t_i) = y_i | \Delta N(t_1) = x_1, \Delta Y(t_1) = y_1, \ldots, \Delta N(t_{i-1}) = x_{i-1}, \Delta Y(t_{i-1}) = y_{i-1})$$

$$\approx \prod_{i=1}^{m} \Phi(t_i)^{x_i} (1 - \Phi(t_i))^{1-x_i} g(x_1, \ldots, x_{m-1}, y_1, \ldots, y_m)$$

Now, assuming m large enough and N simple, $x_i = 1$ i.f.f for some j, $\tau_j \in (x_{i-1}, x_i]$, thus:

$$\sum_{i=1}^{m} (x_i \log(m) + x_i \log(\Delta\Phi(t_i))) - \sum_{j=1}^{n} \log(\phi(\tau_j))$$

$$\sum_{i=1}^{m} \log (1 - \Delta\Phi(t_i))^{1-x_i} \rightarrow - \int_0^1 \phi(t) dt$$

$g(x_1, \ldots, x_{m-1}, y_1, \ldots, y_m)$ some function not involving (the shape of) ϕ. This yields the same results as above.

3.2 Optimality in the sense of least-squares

Another way of viewing the optimality of the fit of a model is optimality in the sense of *least-squares*. In this particular case this could be defined as the integrated squared error, ISE. Traditionally:

$$\begin{aligned} ISE^* &= \int_0^1 (\hat{\phi}(t) - \phi(t))^2 dt \\ &= \int_0^1 \hat{\phi}^2(t) dt - 2 \int_0^1 \hat{\phi}(t)\phi(t) dt + c^* \end{aligned} \tag{3.5}$$

or a weighted version

$$\begin{aligned} ISE^{**} &= \int_0^1 (\hat{\phi}(t) - \phi(t))^2 d\Phi(t) \\ &= \int_0^1 \hat{\phi}^2(t) d\Phi(t) - 2 \int_0^1 \hat{\phi}(t)\phi(t) d\Phi(t) + c^{**} \end{aligned} \tag{3.6}$$

Both versions are not as tractable as the integrated squared error of the $\log(\phi) = \psi$:

$$ISE = \int_0^1 (\hat{\psi}(t) - \psi(t))^2 d\Phi(t) \tag{3.7}$$

$$= \int_0^1 \hat{\psi}^2(t) d\Phi(t) - 2 \int_0^1 \hat{\psi}(t)\psi(t) d\Phi(t) + c \tag{3.8}$$

But equation 3.8 (and equation 3.6), are not empirically computable. On the other hand:

$$\begin{aligned} &\operatorname*{argmax}_{\phi \in \mathcal{S}} \mathcal{L}(\phi, \tau_1, \ldots, \tau_n) \\ &\qquad \equiv \operatorname*{argmax}_{\phi \in \mathcal{S}} \mathcal{L}(\phi, \tau_1, \ldots, \tau_n) - \int_0^1 \log(\phi_0(t)) dt + \int_0^1 \phi_0(t) dt \end{aligned}$$

The following can be derived:

$$\int_0^1 \log\left(\frac{\phi}{\phi_0}\right) d\Phi_0 - \int_0^1 \phi(t)dt + \int_0^1 \phi_0(t)dt =$$

$$\int_0^1 \log\left(1 + \frac{\phi - \phi_0}{\phi_0}\right) d\Phi_0 - \int_0^1 \phi(t)dt + \int_0^1 \phi_0(t)dt$$

using $\log(1+x) \approx x - \frac{1}{2}x^2$

$$\approx -\frac{1}{2}\int_0^1 \left(\frac{\phi - \phi_0}{\phi_0}\right)^2 d\Phi_0 + \int_0^1 \left(\frac{\phi - \phi_0}{\phi_0}\right) d\Phi_0 - \int_0^1 \phi(t)dt + \int_0^1 \phi_0(t)dt =$$

$$= -\frac{1}{2}\int_0^1 \frac{(\phi - \phi_0)^2}{\phi_0} ds \tag{3.9}$$

The integrated squared error of $\log(\phi) = \psi$:

$$\int_0^1 (\log(\phi) - \log(\phi_0))^2 d\Phi_0 =$$

$$\int_0^1 \left(\log(1 + \frac{\phi - \phi_0}{\phi_0})\right)^2 d\Phi_0$$

using $\log(1+x) \approx x$ this time instead of $\log(1+x) \approx x + \frac{1}{2}x^2$

$$\approx \int_0^1 \frac{(\phi - \phi_0)^2}{\phi_0} ds \tag{3.10}$$

Clearly, if ϕ is close enough to ϕ_0 to allow the linear approximation, the maximum likelihood criterion (maximum in equation 3.9) is equivalent to the least-squares minimum as in the version described in equation 3.8 and equation 3.10. Unfortunately, equation 3.8 cannot be evaluated but the maximum likelihood criterion can be used.

3.3 Complexity considerations

The previous two measures assess the deviation between the (estimated) model and 'reality', usually through observed data. It is well known that criteria based on one of these two alone may not suffice. This can be clarified through the following example: suppose we have only one observation. The maximum likelihood method will select a function as peaky as possible. It may be unjustified to assume that points can only occur at that very point, even though that is all information available. This may not be a very realistic example but it serves well as an introduction to the necessity of recognizing the complexity problem.

Complexity can be defined in various ways. The example above shows peakedness as a measure. A common measure is the relative steepness of a function, for instance:

$$\int \left(\frac{\phi'(x)}{\phi(x)}\right)^2 dx$$

or in a weighted version, compare with Good's roughness penalty in § 3.3.2:

$$\int \frac{(\phi'(x))^2}{\phi(x)} dx$$

Another is the maximum periodicity of the function.

There seem to be two ways of restricting the complexity, regardless of the measure used. The most attractive method is the a-priori method in which, based on the observed data and possibly prior knowledge, a certain class of candidate functions is determined. Such a method is discussed in the next section. Some methods attempt to restrict the model while estimating by adding a 'penalty'. Less attractive methods are a-posteriori methods, where a solution is scrutinized after the effort of estimation.

3.3.1 The method of sieves

A method similar to the *method of sieves* Grenander [1981] (as described in [Snyder & Miller, 1990, p 147] or [Karr, 1991, p 229]) can be used to estimate the intensity function. This method is conceived around the use of so-called sieves, sets of functions with particular properties. The properties of the sieves are set in a manner corresponding to the (number of) data-points used. It is hoped that the (constrained) estimated intensity function converges to the true intensity function as the number of points tends to infinity [Snyder & Miller, 1990, p 148].

3.3.2 Penalty methods

An alternative is to penalize the solution for its complexity. A method like the Akaike Information Criterion (AIC) Akaike [1973] could be used. This method is widely used but it focuses on the number of parameters used, not on the shape of the model itself. The AIC is employed as a comparative measure and is computed after (model)estimation. Therefore it does not influence the actual parameter estimates.

Another example is Good's Nonparametric roughness penalty Good & Gaskins [1971], applied in point processes in [Snyder & Miller, 1990, p 151]. This method is based on the Kullback's *information divergence* Kullback [1968] between a distribution and its shifted version. The method of Good & Gaskins [1971] employes the optimization of:

$$\omega(\phi) = \mathcal{L}(\phi) - \mathcal{E}(\phi) \tag{3.11}$$

[Snyder & Miller, 1990, p 147] suggest the use of

$$\mathcal{E}(\phi) = \alpha \int \frac{(\phi'(s))^2}{\phi(s)} ds$$

with α suitably chosen. This results in a certain class of kernel estimators. Study in this direction is useful.

3.3.3 Influence functions and Local-shift sensitivity

From the theory of *influence* functions, [Hampel, Rousseeuw, Ronchetti & Stahel, 1986, chapter 2], the concept of local-shift sensitivity can be used. The

local-shift sensitivity measures the effect of an infinitesimal shift of a point from x to a neighboring point y.

The concept of Local-shift sensitivity can be used to account for the inaccuracy of the observations themselves. Intuitively it is reasonable to restrict the complexity of the model based on observational precision. It is of no use to estimate a periodicity with a wavelength smaller than the error of observation. From a practical point of view it seems a quite reasonable idea to restrict the influence a standard 'deviation' of an observed point can have on an estimate. With respect to an estimate of the intensity at a certain timepoint, the upper bound of the influence over all possible shifts of observations may be a reasonable measure of admissibility of an estimate. In general, the upper bound for all timepoints in the interval may be applicable. Another consideration is what to compare this measure with. It is probably useful to compare this influence with the uncertainty due to the normal estimation procedure.

3.4 Empirical characteristic function

In [Stephens, 1986, par 4.16.5], goodness-of-fit tests are suggested based on characteristic functions. This method can be used in the following manner. Essentially, as suggested in [Stephens, 1986, par 4.16.5], the empirical characteristic function $\gamma_n(s)$ is defined:

$$\gamma_n(s) = n^{-1}\left[\sum_{i=1}^{n} e^{is\tau_i}\right] \tag{3.12}$$

Where $\tau_1, \ldots, \tau_n$ are the sample points. We could attempt to approximate the (assumed continuous) intensity function through an approximate inverse Laplace-transform by:

$$\hat{\phi^*}_{n,T}(x) = \frac{1}{2\pi}\int_{-T}^{T} e^{-isx}\gamma_n(s)d(s)$$

This yields:

$$\begin{aligned}\hat{\phi^*}_{n,T}(t) &= \frac{1}{2n\pi}\sum_{i=1}^{n}\int_{-T}^{T} e^{-isx}e^{is\tau_i}d(s) \\ &= \frac{1}{2n\pi}\sum_{i=1}^{n}\int_{-T}^{T} e^{is(\tau_i - x)}d(s) \\ &= \sum_{i=1}^{n}\frac{\sin(T\ (\tau_i - x))}{n\pi\ (\tau_i - x)}\end{aligned}$$

multiplying by n to get the properly scaled intensity:

$$\hat{\phi}_{n,T}(x) = \sum_{i=1}^{n}\frac{\sin(T\ (\tau_i - x))}{\pi\ (\tau_i - x)} \tag{3.13}$$

$$= \sum_{i=1}^{n} k(T, \tau_i, x) \tag{3.14}$$

In this manner, $\phi(x)$ is estimated through kernels $k(T, \tau_i, x)$. As compared to more common kernels (see Härdle [1990] in a general setting), kernels of this type are smaller than zero for some x, although $\int_{-\infty}^{\infty} k(T, \tau_i, x)dx = 1$.

In [Stephens, 1986, par 4.16.5] the empirical characteristic function is used for goodness-of-fit statistics. Tests similar to Kolmogorov-Smirnov and Cramèr-von Mises are suggested, $\sup_t |\hat{\gamma}(t) - \gamma(t)|$ and $\int |\hat{\gamma}(t) - \gamma(t)|^2 dG(t)$, $G(t)$ suitable chosen, Epps & Pulley [1983]. In a wide sense this is what is done in this work. The gradient test, see § 3.7.3, is -$G(t)$ a counting measure- similar to testing equivalence of

$$\int \sin(nx)\phi(x)dx = \int \sin(nx)dN(x)$$

and

$$\int \cos(nx)\phi(x)dx = \int \cos(nx)dN(x)$$

for a specified value of n.

The question rises whether it is possible to estimate T from the data. More or less this can be rephrased in whether or not the loss-criterion can be minimized.

Traditional maximum-likelihood methods, using $\hat{\phi}_{n,T}$, may not be feasible: if T tends to infinity, the peaks of the kernel, having a value of T/π, dominate the solution (see figure 2) whereas even for bounded intervals I $\int_I k dx \rightarrow 1$ (see figure 3). To put it in other words: as of a certain point the likelihood increases monotonous with T, so T has to be contained in some bounded range.

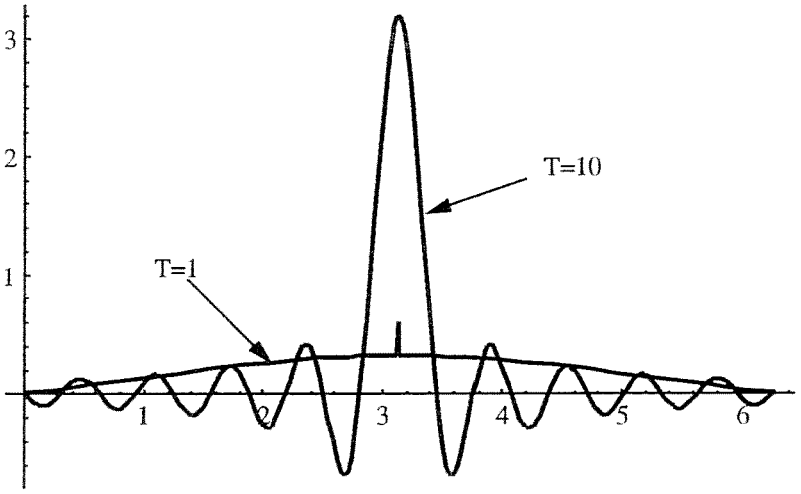

Figure 2: *Empirical characteristic function kernels* $k(T, \tau, x)$ $k(T, \tau, x) = \frac{\sin(T(\tau - x))}{\pi(\tau - x)}$ using $\tau = \pi$, $T = 1, 10$.

It may be possible to estimate T using *cross-validation* along the lines of classical kernel-density estimation techniques like the ones described in [Härdle, 1990, par 4.3]. This technique, using either maximum likelihood cross-validation or least-squares cross-validation, may be employed to get the optimal value of T. Depending on the precise method of cross-validation, this method is less sensitive to the peaks then the previous method. This method has a disadvantage, among others, in that it is computationally of order n^2, n being the number of recorded accidents. Hence it is only a feasible method for small n. What could have been tried is increasing the value of T to the point

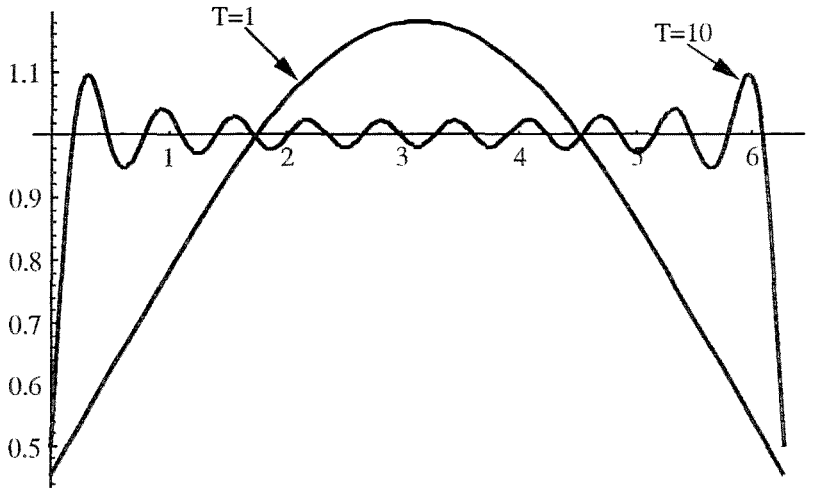

Figure 3: $w(T,x) = \int_0^{2\pi} k(T,\tau,x)d(\tau)$
Effect of limited time-interval. Assuming constant intensity function $\phi \equiv 1$, due to the bounded intervals, estimation on the bounds is biased. In this figure $w(1,x)$ and $w(10,x)$ are drawn. Clearly, $\lim_{T\to\infty} w(T,x) = 1$. In figure 2 it is seen that at the same time $k(T,\tau,x)$ is getting peaked. If both $\phi \to \infty$ and $T \to \infty$, $w(T,x) = \int_0^{2\pi} k(T,\tau,x)\phi(\tau)d(\tau) \to 1$.

where some goodness-of-fit criterion is met. This line of development has been omitted, partly because of the still-large computational effort, partly because of the restricted applicability of kernel-estimates.

Applying the roughness penalty method can also only be done for a small number of points. This can also be stated for all methods based on direct usage of the points. In the next subsection, a method based on a derived statistic is exposed. Methods based on a function of the data, essentially methods based on reduced data, intrinsically employ a kind of restricted model. This will be pointed out § 3.6.

3.5 Assessing model adequacy

Assessing model adequacy is a key step in estimation procedures in general. The choice of methods is determined by (preliminary) assumptions. If a model is assumed to be a member of a particular class, fit of that model to that class can be tested. In general, tests based on such 'subclasses' are more powerful than their more general counterparts. The current case is no exception to that rule. In the followin,g two general tests of fit are discussed. Another possibility, a χ^2-test based on the fit of counts in intervals, is omitted due to the dependence on the construction of those intervals. The method could be used to test sufficiency of a model within a class of models, but may not indicate certain model deviances.

Another problem is that tests have to be carried out while parameters are being estimated. This invalidates many standard procedures or causes considerable loss of power.

3.5.1 *Goodness-of-fit statistics based on spacings*

This method is based on the, rather natural, observation that the time elapsed between accidents bears relation to the model. The idea due to Moran [1951] as explained in Cheng & Stephens [1989] is: Suppose $x_1 < \cdots < x_n$ is an ordered sample of independent stochastic variables with distribution $F_\theta(x)$ where θ is known. Let $y_i = F_\theta(x_i)$:

$$D_i(\theta) = y_i - y_{i-1} \qquad (i = 1, \ldots, m)$$

with $m = n + 1$, $y_0 \equiv 0$ and $y_m \equiv 1$.

$$M(\theta) = -\sum_{i=1}^{m} \log(D_i(\theta))$$

$$\gamma_m = m(\log(m) + \text{Euler } \gamma) - \frac{1}{2} - \frac{1}{12m} + \ldots$$

$$\sigma_m^2 = m(\frac{\pi^2}{6} - 1) - \frac{1}{2} - \frac{1}{6m} + \ldots$$

It is further noted that $M(\theta)$ is asymptotically distributed as $N(\gamma_m, \sigma_m^2)$, although convergence is stated as being slow. Cheng & Stephens [1989] give a small-sample χ^2-approximation.

The authors also give statistics in the case when k parameters are estimated. $M(\hat{\theta})$ is based on $F_{\hat{\theta}}$. The authors define:

$$C_1 = \gamma_m - (\frac{1}{2}n)^{\frac{1}{2}}\sigma_m$$

$$C_2 = (2n)^{-\frac{1}{2}}\sigma_m$$

Then Cheng & Stephens [1989] argue

$$T(\hat{\theta}) = (M(\hat{\theta}) + \frac{1}{2}k - C_1)/C_2$$

is approximately χ_n^2-distributed. $\hat{\theta}$ should be an efficient estimate of θ.

Adaptation to the case where parameters are being estimated is simple and straightforward, as the authors state. This is a major advantage of this method. However, performance in terms of power is somewhere between little and less. Simulations showed that its performance was too mediocre to be useful in practical cases.

3.5.2 *Derived Kolmogorov-Smirnov statistics*

Kolmogorov-Smirnov statistics are often used in practice. The tests are based on comparing the empirical distribution $F_n(x)$ to the hypothetical distribution $F(x)$:

$$D_n = \sup_{-\infty < x < \infty} |F(x) - F_n(x)| \tag{3.15}$$

In the current setting, the fit of $\hat{\theta}(x)$ to $\theta(x)$ is to be tested. Given n and non-stochastic ϕ, the points $\tau_1, \ldots, \tau_n$ are independently distributed with density:

$$f(t) = \frac{\phi(t)}{\int_0^T \phi(s)ds}$$

T being the end point of observation, as was denoted by 1 up to now.

In the standard case, when F is completely specified, confidence limits of D_n are widely available. In the composite case, when parameters are estimated most effort seems be directed to the case where location and scale parameters have been estimated. Most authors denote the general case to be difficult and out of the scope of their work. [Kendall & Stuart, 1987, chapter 30] state that tests are no longer distribution free, but parameter free in the location-scale case. Most authors refer to Durbin [1973]. It seems there is no recent reference about the subject. Even Agostino & Stephens [1986] only offer results in the direction of so called half-sample techniques, for instance Braun [1980] with a method supposedly only valuable in the large sample case. Durbin [1973] offers guidelines for the development of tests, but also mentions a suggestion by Rao [1972], which adapts equation 3.15 in first order approximation about $\hat{\theta}$:

$$\begin{aligned} F_n(x) - F(x) =& F_n(x) - F(x,\hat{\theta}) + (F(x,\theta) - F(x,\hat{\theta})) \\ =& F_n(x) - F(x,\hat{\theta}) + (\theta - \hat{\theta})\frac{\partial F(x,\hat{\theta})}{\partial \hat{\theta}} \end{aligned} \tag{3.16}$$

The key step is the estimation of $(\theta - \hat{\theta})$. Rao [1972] suggested the use of a random sub-sample of the data points containing n' points, about half the sample. From the reference in Durbin [1973] it seems Rao [1972] suggested the use of the first n' points, or they were used for simplicity as is done here. Of course, the points are randomly selected in practice. In the current application each point is selected or not selected based on a (semi)random experiment. This yields a selection of about half the sample. In Rao [1972] probably a selection of exactly half the sample is used, when possible. The estimate of Rao [1972]: ($\mathcal{I}$ is Fisher information of one point)

$$\theta - \hat{\theta} = \frac{2}{n}\sum_{j=1}^{n'} \frac{\partial \log f(x_j,\hat{\theta})'}{\partial \hat{\theta}} \mathcal{I}^{-1}$$

It could be suggested to use $\frac{1}{n'}$ instead of $\frac{2}{n}$ to yield:

$$R_n(x) = F_n(x) - F(x,\hat{\theta}) + \frac{1}{n'}\sum_{j=1}^{n'} \frac{\partial \log f(x_j,\hat{\theta})'}{\partial \hat{\theta}} \mathcal{I}^{-1} \frac{\partial F(x,\hat{\theta})}{\partial \hat{\theta}} \tag{3.17}$$

Using $\sup|R_n|$ instead of D_n in equation 3.15 completes the method. The method has to be adapted to the current situation, considering:

$$F_n(x,\theta) = \frac{\int_0^x \phi(s,\theta)ds}{\int_0^T \phi(s,\theta)ds}$$

This means:

$$\frac{\partial F(x,\theta)}{\partial \theta} = \frac{\int_0^T \phi ds \int_0^x \frac{\partial}{\partial \theta}\phi ds - \int_0^T \frac{\partial}{\partial \theta}\phi ds \int_0^x \phi ds}{\left(\int_0^T \phi ds\right)^2}$$
$$= a(T)b(x) - a(T)^2 b(T)a(x)^{-1} \tag{3.18}$$

with:

$$a(t) = \left(\int_0^t \phi ds\right)^{-1}$$

and the vector

$$b(t) = \left(\int_0^t \frac{\partial}{\partial \theta}\phi ds\right)$$

$$\log f(x,\theta) = \log \phi(x,\theta) - \log \int_0^T \phi ds$$

$$\frac{\partial}{\partial \theta}\log f(x,\theta) = \frac{\partial}{\partial \theta}\log \phi(x,\theta) - \frac{\frac{\partial}{\partial \theta}\int_0^T \phi(x,\theta)dx}{\int_0^T \phi(x,\theta)dx}$$
$$= \frac{\partial}{\partial \theta}\log \phi(x,\theta) - a(T)b(T)$$

defining:

$$c = \sum_{j=1}^{n'} \left(\frac{\partial}{\partial \theta}\log \phi(x,\theta) - a(T)b(T)\right)/n'$$

then, summing:

$$R_n(x) = F_n(x) - F(x,\hat{\theta}) + c'\mathcal{I}^{-1}\left(a(T)b(x) - a(T)^2 b(T)a(x)^{-1}\right) \tag{3.19}$$

Durbin [1973] acknowledges the claim due to Rao [1972] that tests based on $R_n(x)$ behave asymptotically like tests based on $F(x,\theta)$ when $F(x,\theta)$ is completely specified. Clearly, under H_0 if n tends to infinity, c tends to zero, when maximum likelihood is used or any method obtaining an asymptotically equivalent solution. This holds for $F_n(x) - F(x,\hat{\theta})$ as well.

A major disadvantage of this method is the random character of c. This results in the phenomenon that the tests are not replicable. It is clear that this method can be improved by a less arbitrary estimate of $\theta - \hat{\theta}$. Otherwise, the method turns out to be effective and it is employed further.

[Pollard, 1984, Examples 15 and 23] addresses the matter in rigorous manner (example 23) and a less rigorous way in example 15, which may be employed to gain a better approximation.

3.6 Series approximation of functions

3.6.1 Introduction

In practice not very much will be known about the shape of the intensity function. This means a candidate function has to be chosen from a large class (or sieve). At this point there is no obvious choice as to what class of models is suitable. Considering this fact, it seems a good idea to choose a class of models that is relatively easy to use.

The first consideration is about restricting the potential intensity function to be larger than zero at all times or not. This has many consequences, as will become clear later. At first sight, it seems essential to restrict the intensity function to be non-negative. Although post-estimation possibilities may exist, at this point it is opted to constrain the parameters while estimating in order to yield a non-negative intensity function. This can be done, for instance, by actually estimating $\phi = \tilde{\phi}^2$ in the non-negative case or $\phi = \exp(\tilde{\phi})$ in the strict positive case. The latter is pursued here. An advantage of using a quadratic or exponential form is in avoiding serious numerical problems of nonlinear estimation under nonlinear inequality constraints. Both Fletcher [1981] and Luenberger [1984], and probably many more, advise to avoid such constraints when possible. Given the assumption that ϕ can be written as $\exp(\psi)$ for some ψ, a system must be set up to approximate ψ and though this ϕ. A number of choices are available, among them:
– 'short wave' systems: systems consisting of functions that are essentially locally defined. Those functions fade out when moving away from their center. Examples are kernel methods and so-called 'wavelets'.
– 'long wave' systems: polynomial approximation, fourier approximations and the like.
The 'long wave' systems seem to have the advantage of allowing some kind of extension beyond the observed period, commonly called prediction. Because this application might be useful, a combination of 'long wave' systems has been chosen.

When applying a series approximation using terms $f_k(x)$ it is assumed that any continuous function can be written as the infinite linear combination of terms $f_k(x)$.

$$\psi(t) = \sum_{k=0}^{\infty} \theta_k f_k(t)$$

To be practical, all terms $f_k(x)$ should be almost everywhere continuous and it turns out to be virtually necessary that all terms are bounded on a compact set, say between -1 and 1 for reasons shown below. Other practical properties include f being both differentiable and integrable.

The idea [Press, Flannery, Teukalsky & Vetterling, 1989, p.168] is that, given the fact that the coefficients decrease after some index N, 'die out', the deviance is dominated by $|\theta_k f_k(x)| \leq |\theta_k|$. This idea can be found in various other sources as well.

Assumming the intensity function is positive it can be written as:

$$\phi(t) = \exp\left(\sum_{k=0}^{\infty} \theta_k f_k(t)\right)$$

Or as approximation using N terms:

$$\phi_N(t) = \exp\left(\sum_{k=0}^{N} \theta_k f_k(t)\right) \tag{3.20}$$

which should be sufficient, bias is neglected against variance. From now on it is assumed that the intensity function can be defined well by equation 3.20.

It is worthwhile noting that the use of hybrid systems, e.g. combinations of fourier and polynomial systems may have practical advantages. It turned out to be useful that the estimated model has stationary (Fourier) components and a non-stationary (polynomial) component. This fact slightly complicates matter in the following part.

3.6.2 *Loglikelihoodfunction of a series approximation of functions*

This subsection is concerned with the maximum likelihood estimation of functions constructed as above. The loglikelihood is in this case:

$$\begin{aligned}\mathcal{L}(\tau_1, \ldots, \tau_m) &= \sum_{i=1}^{m}\sum_{r=0}^{N} \theta_r f_r(\tau_i) - \int_0^1 \phi_N(t)dt \\ &= \sum_{r=0}^{N} \theta_r \sum_{i=1}^{m} f_r(\tau_i) - \int_0^1 \phi_N(t)dt\end{aligned}$$

Using:

$$C_r = \sum_{i=1}^{m} f_r(\tau_i)$$

this results in:

$$\mathcal{L}(\tau_1, \ldots, \tau_m) = \sum_{r=0}^{N} \theta_r C_r - \int_0^1 \phi_N(t)dt \tag{3.21}$$

It is clear that it is advantageous to have the f_r bounded, so large sums C_r of terms $f_r(\tau_i)$ can be reliably computed from the data. Imagine having to sum a lot of terms t^{100} in the range of $0 < t < \pi$.

Because the terms are bounded and continuously differentiable with respect to θ_k, integrals of $\phi_N(t)$ are differentiable under the integral. This means that the loglikelihood (equation 3.21) is differentiable with respect to θ_k. Its derivative with respect to θ_k is therefore:

$$\frac{\partial \mathcal{L}(\tau_1, \ldots, \tau_m)}{\partial \theta_k} = C_k - \int_0^1 f_k(t)\phi_N(t)dt \tag{3.22}$$

Finally, second order derivatives:

$$\frac{\partial^2 \mathcal{L}(\tau_1, \ldots, \tau_m)}{\partial \theta_k \partial \theta_r} = -\int_0^1 f_k(t) f_r(t)\phi_N(t)dt \tag{3.23}$$

Because $EC_r = E\int_0^1 f_r(t)dN(t)$ and can be estimated by $\sum_{i=1}^{n_i} f_r(\tau_i)$, it follows from theorem 2.4 that the expected value of the gradient is zero if H_0 is true. This result can be used in assessing the relevance of not-used terms, where theorem 2.4 also offers an estimate of its variance.

It has already been noted that in the rest of this work, the set functions consists of the goniometric functions and one trend factor:

$$\begin{aligned} f_{-1}(t) &= t \\ f_0(t) &= 1 \\ f_k(t) &= \begin{cases} \sin\left(\frac{2\pi}{A}\frac{k+1}{2}t\right) & k \text{ odd,} \\ \cos\left(\frac{2\pi}{A}t\right) & k \text{ even} \end{cases} \end{aligned} \tag{3.24}$$

As is obvious, apart from $k = -1$, f_k is bounded on $\mathbb{R}$. At this point it seems useful to indicate a possible confusion. From this point on the meaning of $f_k(t)$ may be ambiguous. When appropriate, it is defined by equation 3.24 and elsewhere it means actually something like $f_{i_k}(t)$, i_k the index in the sense of equation 3.24. It is hoped that this may not result in too much confusion.

In the following subsections asymptotic results are studied, in these cases it is assumed that there is no trend while this would mean that either $\phi(t) \to 0$ or $\phi(t) \to \infty$, which is not considered realistic and for which cases asymptotic results are difficult to obtain, if useful at all. The choice of the f_k, or better, the assumption that the true intensity function can be approximated well by the set functions defined that way, implies some form of stationarity of the true intensity function. This leads to some derived assumptions, using $t \to \infty$:

$$E\left[\frac{\int_0^t \phi(s)ds}{t}\right] \to c \qquad |c| < \infty \tag{3.25}$$

$$E\left[\frac{\int_0^t f_i\phi(s)ds}{t}\right] \to c_i \qquad |c_i| < \infty \tag{3.26}$$

$$E\left[\frac{\int_0^t f_i f_j\phi(s)ds}{t}\right] \to h_{ij} \qquad |h_{ij}| < \infty \tag{3.27}$$

and equivalents to $c(\theta)$, $c_i(\theta)$ and $h_{ij}(\theta)$. These assumptions will be matched with point process equivalences, again, using $t \to \infty$, at least in probability:

$$\begin{aligned} \frac{\int_0^t dN(s)}{t} &\xrightarrow{P} \tilde{c} < \infty \\ \frac{\int_0^t f_i dN(s)}{t} &\xrightarrow{P} \tilde{c}_i \qquad |\tilde{c}_i| < \infty \\ \frac{\int_0^t f_i f_j dN(s)}{t} &\xrightarrow{P} \tilde{h}_{ij} \qquad |\tilde{h}_{ij}| < \infty \end{aligned}$$

When possible, a stronger form of convergence can be assumed, but this is dependent on the specific properties of the point process.

Many authors reduce the parameter space to a compact set, which is quite realistic to do and done here too. The advantages are clear, simplifying many proofs. In this case this practice can be extended to assuming:

$$|c_k| \leq c \leq M < \infty \tag{3.28}$$

M can be taken quite large, say more than the number of accidents that would happen if every inhabitant of the world has a million accidents a day. It is reasonable to assume that serious action would be taken if this number is only remotely approached. However, from a theoretical point of view it is better not to make many assumptions.

3.6.3 *Uniqueness of the solution*

An important question is when the maximum likelihood solution is unique. A heuristic starting point can be found in equation 3.10, from where it can be argued that, assuming asymptotically $ISE \rightarrow 0$, ϕ and ϕ_0 must be equal ds-almost everywhere. This fact combined with a one-to-one relation between θ and $\phi(\theta, t)$, it is likely that, at least in the end, only one θ is the solution. At this point a finite case is dealt with. It is pursued to prove that for every t there exists only one $\hat{\theta}_t$ as the maximizer of the likelihood function.

A first note is that the loglikelihood function is continuous in both C and θ. If C is slightly changed, a new θ can be found close to the previous one. This implies that if there exists a one-to-one mapping from C to θ, this mapping is continuous.

What is shown next is that the Hessian in equation 3.23 is strictly negative definite under certain circumstances. To do this, a combination of matrix algebra and integral convergence theory is used. The crux is that

$$-H_{ij} = (\int_0^1 f_i(t) f_j(t) \phi_N(t) dt)_{ij}$$

can be approximated arbitrarily well by $H(\theta, n)_{ij} = -(I(\theta, n))_{ij}$ in:

$$I(\theta, n) = \frac{\int_0^T \phi(\theta, s) ds}{n} \sum_{k=1}^{n} f_i(x_k(\theta)) f_j(x_k(\theta))$$

using a not equally spaced grid, depending on the distribution Φ. $x_i(\theta)$ is defined in that manner by:

$$x_i(\theta) = \inf_{0 \leq x \leq T} \int_0^x \phi(\theta, s) ds = \frac{i-1}{n-1} \int_0^T \phi(\theta, s) ds$$

This is an adaptation of standard results otherwise found in numerical integration theory. It relies on the assumption that $0 < \phi(\theta, t) < \infty$ for all θ (in the compact set).

On the other hand the matrix $H(\theta, n)$ can be written as

$$H(\theta, n) = -X(\theta, n)' X(\theta, n),$$

$X(\theta, n)$ being the matrix with rows $(f_1(x_i(\theta)), \ldots, f_N(x_i(\theta)))$. Clearly, if the column vectors are independent, $H(\theta, n)$ will be strictly negative definite.

The fact that the columns in $X(\theta, n)$ must be independent for all θ as of some $n > n_0 > 0$, can be translated in the notion that no function f_k can be written as a linear combination of the other functions in all points on the interval $[0, T]$, which would mean that the columns are independent on all $x_i(\theta)$. While θ is assumed in a compact set, the point above can be easily proven.

Another observation is that the limiting Hessian, as defined in equation 3.27, is strictly negative definite too. This is not true in the general case, but it holds by the fact that the f_k are purely periodic.

Clearly, the empirical Hessian $\tilde{H} = -\sum_{i=1}^{N} f_r(\tau_i) f_s(\tau_i)$ may not be definite negative in all, small sample, cases.

3.6.4 *Consistency*

The previous subsection yielded the continuous relation between the vectors C and θ by means of the solution of the maximum likelihood problem.

In this subsection consistency of the estimators is studied. It has already been shown that for every T there is only one maximizing value θ. If the amount of information sequentially increases, the question rises whether there is a limit of the sequence of subsequent estimates of θ, and how it is approached. The key problem is of course the existence problem.

Basically two ways of increasing information are available, one replicating the process, another by increasing the timeframe of observation. Only the last option is practical, but the first option is of theoretical interest. Both options essentially result in increasing the number of observed points.

It is obvious that an important role is played by the point process integrals C, as they determine θ through the likelihood problem. It must be assumed, this cannot be proved by itself, that C_T converges in some sense to a function of T or n. This feature implies some sort of stationarity of the process N_t. In the case of replication, this is obvious, but in the case $T \to \infty$ this requires additional assumptions. Probably the only key assumption is that $N_t < \infty$ if $t < \infty$, but

$$N(t) \to \infty \qquad \text{as} \qquad t \to \infty \tag{3.29}$$

Instead of equation 3.21, a scaled version of the loglikelihood function is used:

$$\mathcal{L}(\tau_1, \ldots, \tau_m) = \left(\sum_{r=0}^{N} \theta_r C_r - \int_0^T \phi_N(t) dt \right) / T$$

Assume, at least in probability (compare equation 3.6.2):

$$C_k(t) = t^{-1} \int_0^t f_k(s) dN(s) \to c_k \quad |c_k| < \infty \quad \forall k \in (0, \ldots, K) \tag{3.30}$$

This means that, given $|c_k| < \infty$ for all $0 < M < \infty$ and $0 < \delta < 1$ there is a $t' > 0$ such that for all $t > t'$ for all k:

$$P(|C_k(t) - c_k| \leq M) \geq 1 - \delta \tag{3.31}$$

Apart from assumptions on the existence of the limit C, [Serfling, 1980, Theorem on page 24], establishes that if f is continuous with P_C-probability one, then $f(C_k) \to f(C)$ accordingly. This theorem cannot be used directly.

Recall that $\hat{\theta}_t = g(t, c(t))$. Convergence of $g(t, y)$ for $t \to \infty$ is studied first. As is seen above, with any probability p, $|y - c| \leq M < \infty$ can be assumed,

thus y in some compact set. This allows to prove only pointwise convergence of $g(t,y) - g(y) \rightarrow 0$ which can be extended to uniform convergence on the set $\{y : |y - c| \leq M\}$.

Pointwise convergence follows from the fact that the limiting Hessian exists and that its eigenvalues are bounded away from zero see § 3.6.3.

Thus, given M and δ small there exists a $t' > 0$ such that for all $\varepsilon > 0$ and $\{y : |y - c| \leq M\}$ $|g(t_1, y) - g(t_2, y)| < \varepsilon$ $t_1, t_2 > t'$ and thus

$$P(|g(t_1, y) - g(t_2, y)| < \varepsilon) \geq 1 - \delta$$

which means

$$g(t_1, y) - g(t_2, y) \xrightarrow{P} 0 \qquad t_1, t_2 \rightarrow \infty$$

This implies, using $(t' < t < k)$:

$$\begin{aligned} |g(t, c(t)) - g(k, c(k))| \leq & \underbrace{|g(t, c(k)) - g(t', c(k))|}_{x_k} + \\ & \underbrace{|g(t, c(t)) - g(t', c(t))|}_{y_t} + \\ & \underbrace{|g(t', c(t)) - g(t', c(k))|}_{z_{tk}} \end{aligned}$$

As seen above, both x_k and y_t converge in probability to 0. By [Serfling, 1980, Theorem on page 24] z_{tk} converges in probability to 0. This proves

$$g(t, c(t)) - g(k, c(k)) \xrightarrow{P} 0 \qquad t, k \rightarrow \infty$$

3.6.5 *Asymptotic normality*

In this section, an attempt is made to estimate the error distribution about θ, assuming the model is *correctly* specified. Treatment in the case of misspecification can be found in White [1982].

As seen above, $\theta_t - \theta \xrightarrow{P} 0, t \rightarrow \infty$. It was assumed that $\lim_{t\rightarrow\infty} \frac{\int_0^t \phi(\theta,s)ds}{t} = p(\theta)$, together with $N(t)/t \xrightarrow{P} c, t \rightarrow \infty$. It was found in § 3.6.3 that C and θ are related, $\hat{\theta}$ is defined as the maximizing value of the loglikelihood function. This means:

$$\frac{\partial}{\partial\theta}\mathcal{L}(\tau_1, \ldots, \tau_m)|_{\theta=\hat{\theta}} = 0$$

This results in:

$$C_k = \int_0^t f_k \phi(\theta, s)ds \qquad \forall k$$

Because θ is estimated consistently, ϕ may well be expanded about its true value:

$$C_k = \int_0^t f_k \exp((\hat{\theta} - \theta)^T f(s))\phi(s)ds \qquad \forall k$$

because

$$\exp((\hat{\theta}-\theta)^T f(s)) \approx 1 + (\hat{\theta}-\theta)^T f(s)$$

it follows

$$C_k \approx \int_0^t f_k \phi(s) ds + \sum_{i=1}^{K} (\hat{\theta}_i - \theta_i) \int_0^t f_k f_i \phi(s) ds \tag{3.32}$$

Define, noting a different notation of H:
- D the vector of $(C_1 - \int_0^t f_1 \phi(s) ds, \ldots, C_K - \int_0^t f_K \phi(s) ds)'$
- $H_{ij} = (\int_0^t f_i f_j \phi(s) ds)_{ij}$,

this results in:

$$(\hat{\theta}-\theta) = H^{-1} D \tag{3.33}$$

Thus

$$E((\hat{\theta}-\theta)(\hat{\theta}-\theta)^T) = H^{-1} E(DD^T) H^{-1}$$

$E(DD^T)$ can be computed through theorem 2.4. It turns out that $E(DD^T) = H$, as defined above.

$$E((\hat{\theta}-\theta)(\hat{\theta}-\theta)^T) = H^{-1} \tag{3.34}$$

After this result, define $\tilde{H} = H/t$ and resume from equation 3.33. Usually, asymptotic normality is achieved by something similar to

$$s(t)(\hat{\theta}_t - \theta) = (tH_t^{-1})(s(t)/t) D_t$$

It is then argued that by:

$$(tH_t^{-1}) \xrightarrow{P} \tilde{H}^{-1} \tag{3.35}$$

and

$$(s(t)/t) D_t \xrightarrow{D} D \approx N(0, \sigma^2) \tag{3.36}$$

that, by Slutski's theorem:

$$s(t)(\hat{\theta}_t - \theta) \xrightarrow{D} \tilde{H}^{-1} D \tag{3.37}$$

It is clear that equation 3.35 follows from equation 3.27. Equation 3.36 is not trivial. Under current assumptions, the martingale D_t converges to a (multivariate) normal distributed vector when suitably scaled by the martingale central limit theorem.

3.7 Algorithm

3.7.1 Overview

In the previous section, the process of estimating a model is exposed. In this section, an algorithm is discussed that uses such estimates to select another, hopefully, better model, hoping that eventually the optimal model is found. This optimal model is supposed to be a member of a certain class, of which members can be selected.

Although in this particular case some sort of fourier system is used, the coefficients of the individual terms cannot be estimated independently. This means

that a model selection scheme needs to be designed to cope with dependencies.

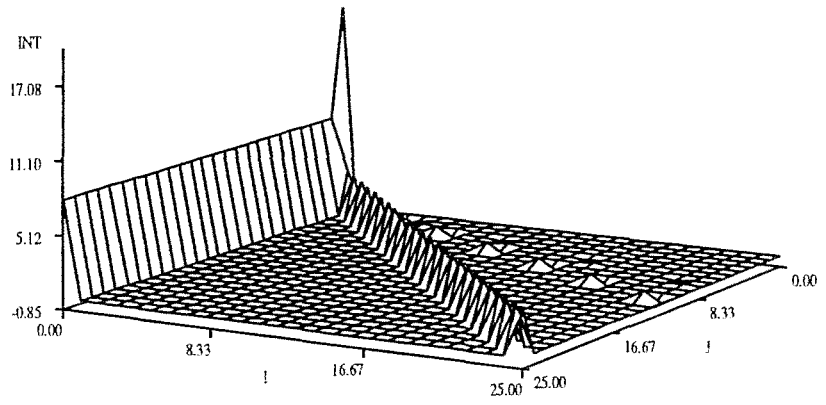

Figure 4: *Integrals of type* $\int f_i exp(f_j)dt$.
For a number of indices the integrals are computed and plotted. Functions $f_i(t)$ are defined as, if i is odd: $f_i(t) = \sin(t(i-1)/2)$, otherwise $f_i(t) = \cos(ti/2)$.

In figure 4 computed values of integrals of $\int f_i \exp(f_j)dt$ are graphed. This is one of the simplest cases where it can be seen that the terms are not independent. It can easily be seen that simple examination of the point process integrals $\sum_i f_r(\tau_i)$ do not offer clear information on relevance of the components. It is clear that estimates of the components will be dependent, as could also be seen from equation 3.23.

The algorithm basically uses three steps:
1 Testing for convergence of the algorithm, not the estimation procedure.
2 Selecting terms to be deleted from the model.
3 Selecting candidate terms to be included in the next estimation procedure.
Other duties of the algorithm include checking if a certain model has already been used.

The procedure is completed when either of the next conditions is met:

The first condition is convergence, achieved when the goodness of fit criterion is met, see § 3.5.2. Usually this point completes the procedure, although sometimes the model can be simplified a little while satisfying goodness of fit. In that case terms in the model are tested for relevance by a method described in § 3.7.2.

Alternatively, the procedure cannot improve the solution anymore. This happens when no suitable terms are left to include. This can be the case because (a) no terms are left at all or (b) all remaining terms are tested insignificant. Two tests for this purpose are described in § 3.7.3 and § 3.7.4.

After a model has been estimated, all terms are sorted by increasing relevance with respect to that model. The terms *in* the model are sorted according to § 3.7.2, the terms not in the model are sorted according to § 3.7.3 or § 3.7.4.

If the algorithm indicates that it has not converged, that is the algorithm indicates that the last estimated model does not satisfy convergence conditions, it first checks whether it can remove any terms. This is done, starting with the least relevant term, until all left over terms test significant. Then (those) terms are added (again) until a model is created that has not yet been estimated. If this means that no terms have been deleted at all, the model adds new terms until either a model is created that has not yet been estimated or all terms tested irrelevant. If the algorithm could not remove any term in the first place, so to say, its guess was completely successful, it can add one or more terms if instructed to do so by a selection scheme described in § 3.7.5.

3.7.2 *Tests for parameters in model*

Suppose one is interested in the relevance of a parameter θ_k in a model. Identify the model by its parameter vector $\hat{\theta}$ and the model, not using θ_k by $\tilde{\theta}$. The relevance of θ_k can be seen comparing the model based on $\hat{\theta}$ to a model based on $\tilde{\theta}$.

Two approaches are commonly used to address this problem:
(1)Estimate $\tilde{\theta}$ by some means and evaluate the likelihood of $\tilde{\theta}$. Compare this value with the original, usually by means of a likelihood ratio test.
(2)Study the surface of the likelihood function about $\hat{\theta}$ and see if it is likely that the likelihood is reduced significantly when the parameter is removed.
Method (1) can be carried out by effectively re-estimating the model, or approximating $\tilde{\theta}$. The latter option turned out too unreliable.

Obviously, method (1) is most reliable but can be very costly to implement in practice. The alternative (2) is implemented through the *Wald*-test (Wald [1943]). Its main attraction is that it is not necessary to estimate the alternative model.

In fact in this problem, only a simple version of the test is necessary. Only the test $\theta_k = 0$ is performed. Generally, define $a(\theta)$ the constraint, define $A = (\partial a(\theta)/\partial\theta_1, \ldots, \partial a(\theta)/\partial\theta_K)$ then:

$$W = -a(\hat{\theta})'(AH^{-1}A')^{-1}_{\hat{\theta}}a(\hat{\theta})$$

This, in practice amounts to:

$$W = \frac{\hat{\theta}_k^2}{H^{-1}_{kk}} \tag{3.38}$$

It is widely known that W is asymptotically χ^2_1 distributed if H_0 is true.

3.7.3 *Gradient test*

The techniques employed in this subsection are closely related to the techniques in § 3.6.5. This test procedure is aimed at testing relevance of parameters not used in the model. In that sense it is similar to so-called *Lagrange-multiplier* tests. It should predict the effect of adding a particular term to the model. The test exposed and used here is designed to add only one term at a time to the model. Lagrange-multiplier tests can test the effects of adding multiple terms to the model.

Suppose the model already contains n terms. These terms each have a certain index i_j, the precise value thereof is not important here. The term k to be tested for inclusion has some real index i_k, but that value is also not relevant in this section. Thus for brevity: $\theta_j \equiv \theta_{i_j}$ and $f_j \equiv f_{i_j}$.

Under H_0 it is assumed that the model is correctly specified. This means that H_0 states that $\theta_k = 0$. Moreover it is assumed that $\hat{\theta}$ is close to the true value of θ. These assumptions should justify the use of the expansions below.

It is assumed that the true intensity function can be written like equation 3.20

$$\phi(s) = \exp\left(\sum_{i=1}^{n} \theta_i f_i(s)\right) \qquad s \in [0, T]$$

The density function $\phi(s)$ will be estimated by $\hat{\phi}(s)$ using the estimate $\hat{\theta}$ of θ. The log likelihood function is (see equation 3.21):

$$\mathcal{L}(\theta) = \theta' C - \int \phi(s) ds$$

in which C denotes the vector of point process integrals.

The maximum likelihood estimator $\hat{\theta}$ is commonly defined as the value of θ that maximizes $\mathcal{L}(\theta)$. The mechanism of the test (and of the Lagrange-multiplier tests) is based on the idea that the loglikelihood can be improved when it has a non-zero derivative with respect to some parameter. Of course, the derivatives with respect to the parameter already in the model are all zero. Thus the derivatives with respect to all non-used terms are evaluated and judged, recall equation 3.22:

$$\left(\frac{\partial \mathcal{L}(\theta)}{\partial \theta_k}\right)_{\theta=\hat{\theta}} = C_k - \int f_k(s)\hat{\phi}(s) ds$$

and equation 3.32:

$$C_k \approx \int_0^t f_k \phi(s) ds + \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i) \int_0^t f_k f_i \phi(s) ds$$

Then, using equation 3.33 and defining

$$v_k = H^{-1}\left(\int_0^t f_k f_1 \phi(s) ds, \ldots, \int_0^t f_k f_n \phi(s) ds\right)',$$

we get:

$$S_k = C_k - \int_0^t f_k \phi(s) ds - v_k' D$$

The vectors D and v_k and the matrix H could be extended by one dimension unit to accommodate k, or the term k represents, to get:

$$w_k = (-v_k \| 1), \qquad R_k = \left(D \| C_k - \int_0^t f_k \phi(s) ds\right)$$

and

$$A_k = E\left[R_k R_k^T\right]$$

This results in:

$$S_k = w_k' R_k \tag{3.39}$$

Under H_0, S_k will be approximately normal distributed with mean 0 and variance $w_k' A_k w_k$. The expression $w_k' A_k w_k$ can be simplified further.

The alternative is the Lagrange-multiplier test. This test needs a (possibly large) matrix inversion for every tested parameter. It then amounts to, using $r_k = C_k - \int_0^t f_k \hat{\phi}(s) ds$:

$$LM_k = r_k^2 (A_k^{-1})_{n+1,n+1}$$

Although the Lagrange-multiplier test may be a useful test, it is not used because it also needs computation of the integral and it needs a sometimes time-consuming matrix inversion. Another disadvantage is, albeit only to some extent, that the method is incompatible with the method in § 3.7.4. Some computational relief can be found by applying the *Sherman-Morrison* formula [Press, Flannery, Teukalsky & Vetterling, 1989, p.75]. This formula can be used to invert a matrix when the inverse of a similar matrix is known. This similar matrix being the covariance matrix of the parameters in the model.

3.7.4 Approximate gradient test

The derivation of the approximate test needs a little explanation. The main reason for using an approximate test instead of a less approximate test as in § 3.7.3 will be determined by computational considerations. The reason for developing this test is the sometimes herculean effort needed to compute a (very) large number of strongly periodical numerical integrals. This can be *very* time-consuming. This case arises when a large number of accidents is analysed over a number of years that have a strong daily pattern. Although this kind of problem may not occur very often, some preparations can be made in advance.

A few options are available to lighten the computational burden. First and foremost, restricting the number of times all integrals have to be computed. This is done to some extent by a step described in § 3.7.5, where a (heuristic) strategy is exposed that 'makes larger steps' in the selection scheme. Of course it is hoped that those steps don't include too much terms that have to be rejected at second sight, which can be a major drawback.

Another option is to reduce the number of integrals that have to be evaluated. This method is based on partial integration and exploiting properties of trigonometric functions. The trick is to expand a function in two different functions in an practical way. For instance, $f(x) = g(x) + h(x)$ then $\int f(x)dx = \int g(x)dx + \int h(x)dx$. Doing this in a clever manner, having computed $\int g(x)dx$ already means by computing $\int f(x)dx$ we get $\int h(x)dx$(for free). More details follow.

It is assumed that the functions $f_i, i \geq 0$ obey the following rules:
- f_0 is constant.
- $f_i(s)$ is continuous and differentiable with respect to s on $[0, T]$.
- The set of functions $\mathcal{T} = \{f_i | i = 1, 2, \ldots\} \cup \{f_0\}$ is closed under products, that is, $f_i f_j$ can be written as a linear combination of elements of $\mathcal{T}$.
- with respect to s, $\mathcal{T}$ is closed under both differentiation and integration. Consequently, f_i is infinitely differentiable and integrable on $[0, T]$.

The set of functions used here obeys these rules. Recall:

$$\phi'(t) = \sum_{i=1}^{n} \theta_i f_i'(t)\phi(t)$$

For $k > 0$, if the right model is specified:

$$\int_0^t f_k \phi ds = [F_k \phi]_0^t - \sum_{i=1}^{n} \theta_i \int_0^t F_k f_i' \phi ds \tag{3.40}$$

This relation is to be exploited in this subsection. Alternatively, the relation holds for the estimated case. It doesn't rely on the model being correctly specified.

$$\int_0^t f_k \hat{\phi} ds = [F_k \hat{\phi}]_0^t - \sum_{i=1}^{n} \hat{\theta}_i \int_0^t F_k f_i' \hat{\phi} ds \tag{3.41}$$

It is noted that the latter kind of integrals: $\int_0^t F_k f_i' \hat{\phi} ds$ can be expanded into two integrals of type $\alpha \int_0^t f_{i_\alpha} \hat{\phi} ds$ and $\beta \int_0^t f_{i_\beta} \hat{\phi} ds$. The main object of the test is to find out whether or not $\int f_k dN \approx \int f_k \hat{\phi} ds$. Under H_0, the model is correctly specified, this is the case. The idea is to simply put $\int * dN$ in place of $\int * \hat{\phi} ds$ in equation 3.41 and test equivalence of:

$$\int_0^t f_k dN(s) = [F_k \hat{\phi}]_0^t - \sum_{i=1}^{n} \hat{\theta}_i \int_0^t F_k f_i' dN(s) \tag{3.42}$$

as $\hat{\theta}_i$ and $\int_0^t F_k f_i' dN(s)$ are not independent, a test based on this is biased. Equation 3.42 and equation 3.41 are not even asymptotically equivalent. It is assumed that the bias is negligible with respect to the variance. Moreover, this test may not be used to test parameters in the final stages. It may serve as an intermediate to speed things up.

With respect to equation 3.41, it can be seen that $\int_0^t f_k \hat{\phi} ds$ can be expressed as a combination of other integrals. It seems attractive to select the most complex integral on the right side of equation 3.41 and solve for this equation by computing the others. This scheme enables the (recursive) computation of all integrals from some starting point. This starting point is at most right above the most complex term used in the model. A few notes however:
– Fortunately, this is a numerical scheme so stochastic properties of $\hat{\theta}$ are of no influence.
– Unfortunately, the method relies strongly on the precision at which the integrals are computed.
– The method only works for the terms with complexity larger than the maximum complexity in the (estimated) model. This means that if the maximum number of terms is estimated well in advance, this number will not be much higher then the maximum complexity in the model. This in turn means that the advantage is only based on not having to compute integrals for a relatively small number of terms.
– If the parameter θ_n of the most complex term in the model is rather small, this also adds to the numerical instability of the procedure.
The above sketched notes help to establish the conclusion that the recursive method may not be reliable in general and it is skipped therefore. Computational advantages are doubtful too.

On the other hand, a test based on equation 3.42 appears to be useful. The test is not used without the backing of § 3.7.3. 'Critical' terms are tested by § 3.7.3. The test statistic is defined as:

$$\tilde{S}_k = \int_0^t f_k dN(s) - [F_k \hat{\phi}]_0^t + \sum_{i=1}^{n} \hat{\theta}_i \int_0^t F_k f_i' dN(s) \tag{3.43}$$

It is assumed that the distribution of $\tilde{S}_k$ has a mean value of approximately nil. It is also assumed that $|\tilde{S}_k|/\tilde{\sigma}^2(\tilde{S}_k)$ reflects the order of the likelihood under H_0 of the individual terms being nil. $\tilde{\sigma}^2(\tilde{S}_k)$ can be computed assuming a multidimensional normal distribution. In order to compute the covariances involving $\hat{\theta}$ equation 3.34 is used to express terms of $\hat{\theta}_i$ in terms of simple point process integrals. The terms $\int_0^t F_k f_i' dN(s)$ are also expanded into simple point process integrals. Then theorem 2.4 is used to compute the variance. It turns out that second order terms can be neglected:

$$\begin{aligned}\sigma^2(xy) &= \mu_x^2\sigma^2(y) + \mu_y^2\sigma^2(x) + 2\mu_x\mu_y cov(x,y) \\ &\quad * cov^2(x,y) + cov(x,x)cov(y,y) \\ &\approx \mu_x^2\sigma^2(y) + \mu_y^2\sigma^2(x) + 2\mu_x\mu_y cov(x,y)\end{aligned}$$

The above derived method won't work for f_0 and f_{-1}. Integrals for these functions will have to be computed in the classical manner.

3.7.5 Selecting more terms at a time

The algorithm defined so far is based on the idea that only one new term at a time is included in the model. Then while systematically checking the usefulness of terms already in the model, terms are added as long as that seems necessary, until convergence is met. When a complex system is to be estimated, this means that quite a few steps must be made, each needing a full gradient test and a full goodness-of-fit test. It seems useful to add a few controlled leaps to this process, in order to speed things up. This subsection describes a way of doing this.

This method is based on the idea that the Hessian of the loglikelihood should not have too different eigenvalues. In more practical language this means there should be little dependence between the terms in the model. The implementation is again based on the idea that $\int f_i f_j \hat{\phi} ds \approx \int f_i f_j dN(s)$ when the model is correctly specified.

The method is used after the parameters have been selected by a combination of either § 3.7.3 or § 3.7.4. This defines a queue of terms of which step by step a term is added until one of the following events occur:
– No terms are available or the maximum number of terms to be added is overdrawn. This parameter should be user supplied.
– The next term did not test significant in the sense of § 3.7.3. The terms that could possibly be included in the model are always tested by § 3.7.3. These terms are among the 'critical' terms mentioned in § 3.7.4.
This step supplies an ordered range of terms from which terms can be selected. While waiting for the above mentioned event to occur, the procedure computes a relative effect of adding up to a particular term to the model. This effect is the *condition* of the Hessian, which is the quotient of the largest and the smallest eigenvalues of the Hessian. This value could be computed by computing (part

of) the eigenvalue spectrum. This method can be costly in some cases. As the method is not needed to work perfectly, the condition is estimated by estimating the largest and the smallest eigenvalue of the empirical Hessian using the theorem of Gershgorin, [Stoer & Bulirsch, 1980, p. 385]: The union of all disks

$$K_i = \left\{ \mu \in \mathbb{C} \mid |\mu - a_{ii}| \leq \sum_{k \neq i} |a_{ik}| \right\}$$

contains all eigenvalues of the $n \times n$ matrix $A = [a_{ij}]$. Thus the maximum and minimum in $\mathbb{R}$ of this set are used.

While adding at least one term, it is hoped that the number of terms just before the greatest jump is a good candidate to use in the next algorithmic step.

4. Extensions

As it is defined so far, the model may not be very useful in practice. Some extensions have to be made in order to get some information from an analysis. Uses of data analysis generally are in one of two categories:
(1)Retrospective analysis. Roughly, finding out what happened after it happened
(2)Prospective analysis. Roughly, what is likely to happen in the future.
Both types, but mainly (2), are deeply buried in assumptions. Case (1) could be divided into a pure phenomenalistic part and an exploratory part. The first is to find out what the intensity was at a certain timepoint, no matter what caused it to be that way.

More often one will try to explain the nature of the process using exogenous variables. See § 4.1 for some remarks on the use of exogenous variables.

Probably the most interesting application is prediction. Prediction is not meant in the sense of predicting the times accidents occur. The main interest will be the expected number of accidents in a time period. Another may be the conclusion that the number of accidents of a certain kind seems to rise or not.

The prediction above is meant in a literal sense. Another application of prediction schemes is 'prediction of the past'. This is explained in conjunction with intervention analysis in § 4.3. Prediction in general is a subject in § 4.2

4.1 Exogenous variables

The inclusion of exogenous variables seems a quite straightforward job. It is best to note that exogenous variables could be regarded as simple functions x_t of the time. Only a few remarks have to be made. As a consequence of § 3.3.3, it seems highly advisable to assume the functions induced by exogenous variables to be continuous in t. This would keep ϕ continuous. In most cases this may not be an unreasonable assumption anyway. Its main advantage will be less sensitivity to local aberrations, both in the choice of the change points of the external function and in the accident points around the same point. This is not a compelling advise though.

Relevance of exogenous variables that are not included in the model can be tested through § 3.7.3 but in general not through § 3.7.4. Relevance of exogenous variables that are in the model can be tested through the *Wald*-test, see § 3.7.2. No exogenous variables have been used here. It may be useful to give (some) exogenous variables a special role in the model selection process. In that case they should be exempt from removal of the model. If one is interested in a comparison of solutions with and without certain exogenous variables, it seems attractive to have control over their inclusion in the model.

Another (small) advantage of the continuity of the functions defined by the exogenous variables is that their integrals with respect to ϕ are more easily computed.

Suppose $x(t)$ has points of discontinuity in $x_1, \ldots, x_r$, then:

$$\int x(s)\phi(s)ds$$
$$= \sum_{i=1}^{r} \left(\int_{x_{i-1}}^{x_i} x(s)\phi(s)ds + \phi(x_i) \left(\lim_{s \downarrow x_i} x(s) - \lim_{s \uparrow x_i} x(s) \right) \right)$$

This of course is not a serious problem, but it can complicate things.

A completely different point of view on exogenous variables is to compare the estimated intensity function to some exogenous variables. This case seems a little out of scope here, so it is omitted. A particular application of exogenous variables is highlighted in section § 4.3.

4.2 Some derived statistics

A number of statistics can be derived from a solution of the model. The main statistic will be the solution vector and its covariance matrix, as the solution vector is assumed to be normally distributed. Most statistics, if not all, will be derived from these.

4.2.1 The intensity at time t

An obvious choice. Technically this could be divided into prediction and filtering, computation outside the interval of estimation or computation inside the interval of estimation. It has to be stressed that great care must be taken using models to predict behaviour outside their interval of estimation, but usually it is the only available option, particularly if one is dealing with the future. As compared to prediction methods using transfer functions, as is common in timeseries modelling, no information is available on the decrease in prediction quality as the (lead) time increases. This may be a serious shortcoming of the method in this context. The confidence interval of the prediction is only based on the distribution of the parameters, not on the time elapsed. This may not be a big problem if a short period ahead is predicted. The rest is simple. Assume H is the covariance matrix of $\hat{\theta}$, the exponent is computed as

$$\hat{\psi}(t) = \sum_{i=1}^{n} \hat{\theta}_i f_i(t) \qquad \hat{\phi}(t) = \exp(\hat{\psi}(t))$$

and $\sigma^2(\hat{\psi}(t)) = \sum_{i=1}^{n} \sum_{j=1}^{n} f_i(t) f_j(t) H_{ij}$ Using this result a confidence interval can be computed, using normality assumptions. This results in $\psi(t) \in [\psi_l(t), \psi_u(t)]$. It is then assumed that $\phi(t) \in [\exp(\psi_l(t)), \exp(\psi_u(t))]$.

4.2.2 The cumulated residuals at time t

Another application is the equivalent of a *cusum* (cumulative sum) analysis in timeseries analysis or quality control. In this context it is a comparison of observed points to the predictions of the model. Essentially this is the cumulated residual of the model up to time t:

$$CR(t) = N(t) - \int_0^t \hat{\phi}(s)ds \quad t \geq t_0 \tag{4.1}$$

This function can be defined both within the timeframe of estimation and outside the timeframe of estimation. The latter will be the usual application.

The version within the timeframe of estimation is usually used in the context of goodness-of-fit analysis, see § 3.5.2 for more details. The alternative is usually used to test model validity to support predictions. Technically, referring to cumulative sum techniques, the cumulative residual technique is used to find out if or when the data diverges from the model. Ideally it is found that the model seems to fit the data well after the estimation interval, suggesting the model might also do so in the future. Other applications include intervention analysis in § 4.3.

It is assumed that the accidents after the timeframe are independent (or its dependence can be neglected) of the accidents within the timeframe, based on which the parameters are estimated. Therefore it is assumed that the error distributions in $N(t) - N(t_0)$ and $\hat{\phi}$ are independent. Then, using the usual normality assumptions, a confidence interval for $CR(t)$ in equation 4.1 can be computed. Under H_0, the model being correctly specified *and* valid after estimation period: $t_0 > T$, T the end of the estimation period.

$$\sigma^2(CR(t)) = \sigma^2(N(t) - N(t_0)) + \sigma^2(\int_{t_0}^{t} \hat{\phi}(s)ds)$$

Then it is (only) assumed that $\sigma^2(N(t) - N(t_0)) = \int_{t_0}^{t} \hat{\phi}(s)ds$ and that $\sigma^2(\int_{t_0}^{t} \hat{\phi}(s)ds)$ can be well approximated to first order by:

$$\sigma^2(\int_{t_0}^{t} \hat{\phi}(s)ds) = g'Hg$$

H is the covariance matrix of $\hat{\theta}$ and $g = (\int_{t_0}^{t}(\partial\hat{\phi}(s)/\partial\theta)ds$. One problem that is not solved is that $\sigma^2(N(t)-N(t_0)) = \int_{t_0}^{t} \hat{\phi}(s)ds$ may not be a very good approximation because something similar to *overdispersion* or *underdispersion* might occur. This will influence the results, by systematically estimating the variance incorrectly.

4.2.3 *The integrated intensity at time t*

The integrated intensity is the integral of the intensity alone. It can be viewed as a combination of § 4.2.1 and § 4.2.2. The main difference will be that it is computed for intervals. Thus:

$$\text{I-Int}(t_k) = \int_{t_{k-1}}^{t_k} \hat{\phi}(s)ds \tag{4.2}$$

Computation of confidence intervals is almost equivalent to the case of § 4.2.2, except for the fact that the variance of separate intervals may not add up to the variance of the joined interval. This can be seen by:

$$(g_1 + g_2)'H(g_1 + g_2) \neq g_1'Hg_1 + g_2'Hg_2$$

The difference $2g_1'Hg_2$ may not cancel *outside* the interval of estimation.

4.3 Intervention analysis

As already mentioned, intervention analysis is an application of this method. As of Box & Tiao [1975] and earlier, a large number of applications have been proposed, including Harvey [1985], with application to road safety research Harvey [1986] and using the same method Ernst & Brüning [1990] also in road safety research. Although these methods are based on different assumptions, a few of these viewpoints can be used in this context:

– An intervention can cause more that just a change in level. In applied work one should not lose the main object under investigation. In most traffic safety work the prime interest is (the number of) accidents or casualties. Sometimes it can be useful to study a change in the pattern, not just the sheer number. Clearly the prime interest is some function of the process.

– The use of control groups in experimental research is essential: a change found in some process should always be supported by not finding a similar change in another (relevant) process. This suggests the use of multivariate models, analysing more processes at one time.

– Multivariate models are difficult to handle and good implementations seem rare to find. [Harvey, 1985, p 39] states that, under certain (homogeneity) conditions the single dimensional (single equation in terms of Harvey [1985]) model might suffice. It seems more recent work (f.i Kendall & Ord [1990]) hardly improves this situation. This means that experimental groups and control groups must be analysed separately in most cases, our case is no exception.

– Apart from time series technicalities, it seems cardinal to identify interventions first 'by them selves', for instance using a cumulative sum technique, used by Harvey [1986] but not by Box & Tiao [1975]. This technique works as follows: assume an intervention is supposed to take place at time t_1. Then a time series is estimated (or identified) up to a time point t_0 well before t_1. What is meant by 'well before' will be the ever returning expert's guess. Then the estimated model is used to predict the observations as of t_0. This is analysed by the cumulative sum technique. Hopefully, the model seems to fit for some time after t_0, supporting the suggestion that the model is correctly identified. If the intervention really had an effect, it is assumed that the model diverges, or better, misfits, after or about t_1. This method should be favored over methods that simply use plug-in type dummy variables representing interventions, and validations of those interventions based on the importance of those dummy variables in those models. Not seldom one seems to find a significant effect using a dummy variable while the true intervention point is at some other time and the change in the series is due to something else. Therefore it is regarded as favorable to estimate the intervention time before modeling it. Of course, the estimate of the intervention time may differ a little from the intervention time that is to be modeled, but it may not be too far off to allow for alternative explanations. If the intervention is positively identified, then (possibly dummy) exogenous methods can be used. One problem of the above mentioned method is clearly the availability of sufficient data, particularly before the intervention point. This is a particular problem in the area of traffic safety where sometimes interventions are executed shortly after a problem is identified or suitably measured. It is a rare case where a problem is identified and studied for a number of years before some intervention is made. It usually results in non-systematic information (changing over time) and statistical problems thereof. At this point the model is restricted to the search of interventions points only. Both the cumulative sum and the 'before-after' techniques are supported, al-

though the cumulative sum technique is emphasized. Implementation of the cumulative sum technique can be found in § 4.2.2.

In Chapter 5 and Chapter 6 examples are shown.

5. The simulation of problems

5.1 Overview

Obviously, introduction of a new method or even a new implementation requires careful testing. In this section results of some testing are shown. Hopefully it covers a sufficiently broad spectrum of problems.

As mentioned in various previous sections, the intensity function is assumed to be an element of a certain class. This class is taken to be like equation 3.24. A predominant class of tests is based on that assumption. Another situation is when the true intensity is *not* in that class. From another point of view, stochastic properties can be studied. These consist of asymptotic results, overdispersion and the like. The first is done here, albeit approximately, the second is not. It is omitted mainly to reduce the amount of work involved and the necessity of adapting the simulation process. What has been done in this direction is a little degenerated, removing all stochastics in the simulation. This was done in the early stages and not reported here.

Apart from testing the estimation process, the derived statistics can be studied too. Particularly the cumulative sum techniques need attention. Results can be found in § 5.4.

5.2 Simulating accident data

In this subsection a strategy is shown for simulating data. It is mainly based on [Grandell, 1990, lemma 4 and lemma 5, page 34]. This lemma states that non-homogenous Poisson processes can be converted in homogenous Poisson processes and vice-versa. This is done through a time transformation.

The procedure is now as follows:
(1)Generate an intensity function that has the required properties.
(2)Define the interval of estimation, denote it by $[t_0, T]$.
(3)Produce exponentially distributed random data points $\delta_1, \delta_2, \ldots$ and:
(4)Compute t_k as the solution of:

$$\int_{t_{k-1}}^{t_k} \phi(s)ds = \delta_k \tag{5.1}$$

(5)Stop if the number of points reaches a specified maximum or if $T < t_{n+1}$. Then n points have been simulated.
Alongside this simulation a number of statistics are being computed. These are merely designed to relieve the experimenter when the model fails to fit the data in a sufficient manner by testing the true intensity by a Kolmogorov-Smirnov test or when certain properties could not be recovered from the analysis. For these cases the internal Wald test (§ 3.7.2) is computed to test whether the true parameters could be identified at all from the data. These tests can fail to give the expected results because of the random nature of the experiments *and*, probably most important, because not enough data is used to identify the model properly.

Concluding, only Poisson models have been used to test here. Apart from exponential intensity functions, quadratic models have been used.

5.3 Practical problems

A few practical problems are noteworthy.
– Equation 5.1 cannot be solved analytically. This could have been solved by a numerical method but this was deemed (again) too computationally intensive in case many points are simulated. It was chosen to approximate the intensity function by a stepfunction. This relieved the previous problem but it introduced the possibility of the *Gibbs*-phenomenon, well known from spectral analysis (unexpected peaks due to the 'sharp edges' created by the stepfunction). It should be borne in mind that this effect can occur. It also restricted the maximum complexity of the simulations somewhat.
– The expected number of simulated points is an attractive feature to have control over. This means that, given randomly chosen parameters θ_i, it effectively modifies the generated parameters by forcing $\int \phi ds = n$ for some n. This in turn results in a not independent generation of the parameters. This last effect has been neglected.
– In the simulations, the number of expected points is controlled. The random simulated parameter values generally do not result in an intensity with the expected number of points. A correction scheme is designed for this. The correction scheme of the expected number of points is different for exponential models and quadratic models, although the mathematical mechanism is the same. Simply ϕ is multiplied by a constant. In the exponential case this merely results in changing θ_0, in the quadratic case all parameters are multiplied. Adding a constant to θ_0 in this case leads to relative flattening of the resulting intensity function, yielding unidentifiable parameters, which was found unwanted.
Alongside the points, a dataset is created containing the true intensity and an estimate of it based on the generated points using the kernel-method of § 3.4. The choice of the kernel width was based on the maximum complexity of the simulation. This number was thus known ahead, which does not reflect reality. The results can be seen as indicative, and can be used to compare with the estimated intensity generated by the maximum likelihood method.

5.4 Some examples

5.4.1 $n \to \infty$ *and more*

In this simulation an intensity function consisting of five terms, level, $\cos(\pi t)$, $\cos(2\pi t)$, $\cos(4\pi t)$ and $\sin(4\pi t)$ is used. No trend was included. The simulation ranged from 0 to 4, thus containing two entire cycles. The models were estimated on the data of the first period (0 to 2). A total number of points of 1000, 3000, 5000 and 7000 were anticipated for the total period, so about half of them were used in actual estimations. This is the most important reason for not including a trend in the model.

The terms, and some results are listed in table 1. The level term was omitted in table 1 because its value is different depending on the number of points. The first model ($n = 1000$) failed to identify the parameters correctly using the standard scheme. It did identify the parameters when the analysis was extended until also the gradient test was satisfied. This method is now called the extended model.

Coefficient	n=1000	n=3000	n=5000	n=7000
$\cos(\pi\, t)$	-0.769	-0.6097	-0.6026	-0.6383
$\cos(2\pi\, t)$	0.000	0.5425	0.6159	0.5639
$\cos(4\pi\, t)$	0.000	-0.2203	-0.2245	-0.2101
$\sin(4\pi\, t)$	0.000	-0.1906	-0.2043	-0.2197

Table 1: *Subsequent values of the parameter estimates while increasing n.*

The following figures show the prediction (or restoration) of the intensity function. In figure 5 ($n = 1000$), figure 20 ($n = 3000$),figure 21 ($n = 5000$) and figure 22 ($n = 7000$), the (95%) upper and lower limit are graphed together with the true intensity function. It is visible that the intensity is not perfectly contained in the confidence region in figure 5. This graph uses the extended model. It may be the case that the model is not functioning very well at this number of points or the confidence ranges may be too small, because the deviance is not that much so it seems. In the other graphs, no such problems occur. These graphs, figure 20 ($n = 3000$), figure 21 ($n = 5000$) and figure 22 ($n = 7000$) are listed in the appendix.

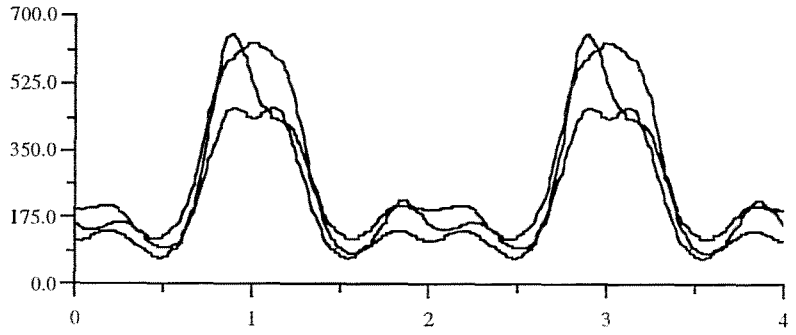

Figure 5: *Predictions and true intensity of $n = 1000$.*
In contrast to the models based on more points, this model seems to show an insufficient confidence region. This problem sometimes arises when few points are used. Of course, this can also be caused by the random nature of the model.

Another feature is the cumulative residual, see § 4.2.2. In the following the figures are figure 6 ($n = 1000$), figure 7 ($n = 1000$) and figure 8 ($n = 7000$), figure 23 ($n = 3000$), figure 24 ($n = 5000$) are in the appendix. The most striking effect may be the fact that in figure 8 ($n = 7000$) the cumulative residuals indicate a diversion from the model. This effect also surfaced in the $n = 10000$-case (which is further omitted). In general, a tendency toward 'breaking out' through the lower limit seems to show up. This may be an indication of a non-symmetrical distribution, which is not assumed in § 4.2.2 where normality is assumed. This phenomenon has not been studied further, although there might be sufficient reason to do so.

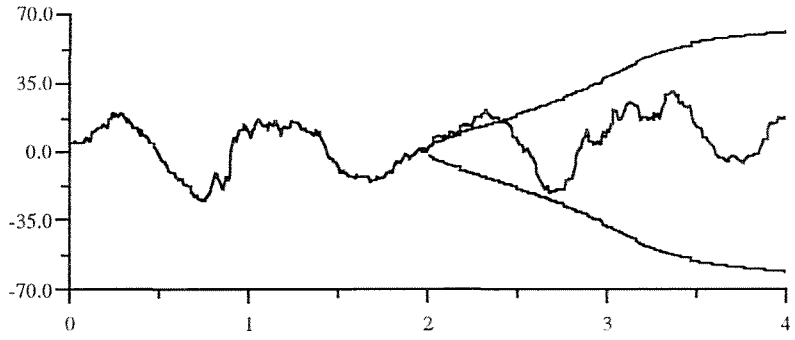

Figure 6: *Cumulative residual graph of* $n = 1000$.
Note the relatively large fluctuations in the residuals. This is possibly due to premature convergence. The estimation procedure was extended to satisfy integral tests as well. The results are graphed in figure 7.

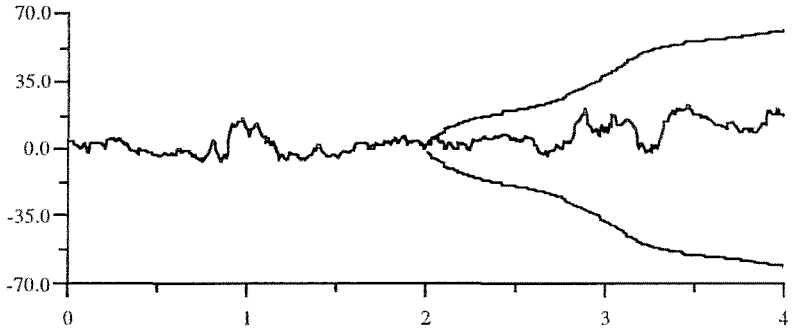

Figure 7: *Cumulative residual graph of* $n = 1000$. *Integral convergence*
The model used to create these results satisfied both goodness of fit and the gradient test of § 3.7.3. Clearly the results are better. Improvement of this kind is rare however, in most cases the both results are identical. This may indicate that the goodness of fit tests needs additional support in small-sample cases, which is not surprising.

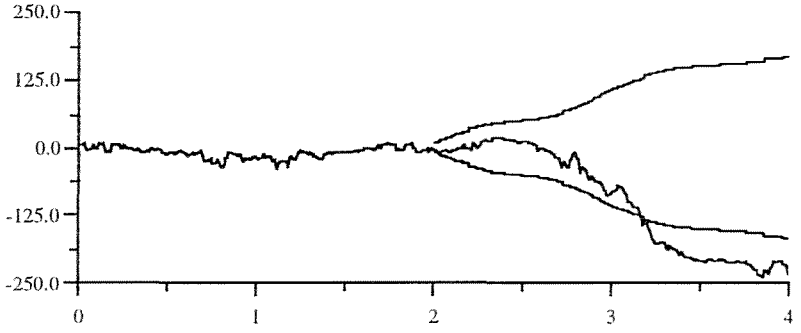

Figure 8: *Cumulative residual graph of* $n = 7000$.

Another explanation may be that the prediction 'lead' is simply too long. Although it sometimes seems to work pretty well, it may be risky to predict that long ahead compared to the length of the estimation period. Extending the estimation like in the $n = 1000$ case did not solve this problem.

Yet another feature is the integrated predictions based on the integrated intensity, see § 4.2.3. Using a more or less arbitrary 0.01 interval length, the number of points in those intervals are graphed together with the confidence intervals based on § 4.2.3. Although the method of computation is similar to the computation of the cumulative residues, no indication seems to show up that the lower limits seem to be to high. This supports the idea that the predictive period is simply too long to be reliable in figure 8, although individual counts seem to fit sufficiently.

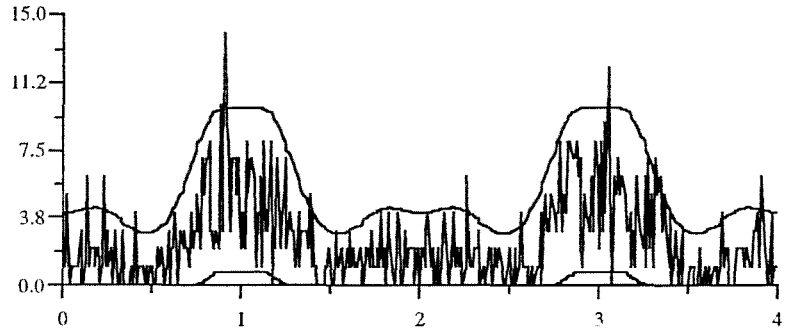

Figure 9: *Graph of integrated prediction interval and tabulated points* $n = 1000$, *extended estimation.*

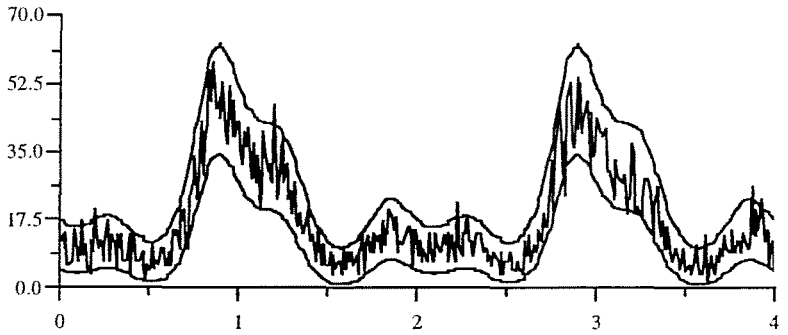

Figure 10: *Graph of integrated prediction interval and tabulated points* $n = 7000$.

The procedure is the same as for the predictions and the residues above. The case ($n = 1000$, with extension) and ($n = 7000$) are graphed here in figure 9 ($n = 1000$) and figure 10. The others, figure 25 ($n = 3000$) and figure 26 ($n = 5000$) are in the appendix. Only rarely the number of observed points is outside the confidence interval.

5.4.2 *Indicating underspecification*

In this set of simulations a term is modified to a specific order. The object of this simulation is to try identifying the case that the model is underspecified: the maximum complexity is too small. The anticipated condition is that the gradient test (§ 3.7.3) indicates the model cannot be improved while the (modified) Kolmogorov-Smirnov test (§ 3.5.2) indicates lack of fit.

Term	θ
Trend	0.07870
$\cos(\pi\, t)^*$	0.04918
$\sin(2\pi\, t)$	0.66789
$\sin(3\pi\, t)$	-0.37012
$\cos(3\pi\, t)$	0.48918
$\cos(20\pi\, t)$	1.00000

Table 2: *Parameters and terms used to test underspecification*
$^*\cos(\pi t)$ did not test significant in the $n = 1000$ case. The wald test based on the true parameter value was significant at the 3.03484×10^{-1} level. The model did not identify the term.

In this case the new term is $\cos(20\pi t)$. The other terms are: trend, level, $\cos(\pi t)$, $\sin(2\pi t)$, $\sin(3\pi t)$ and $\cos(3\pi t)$, see table 2.

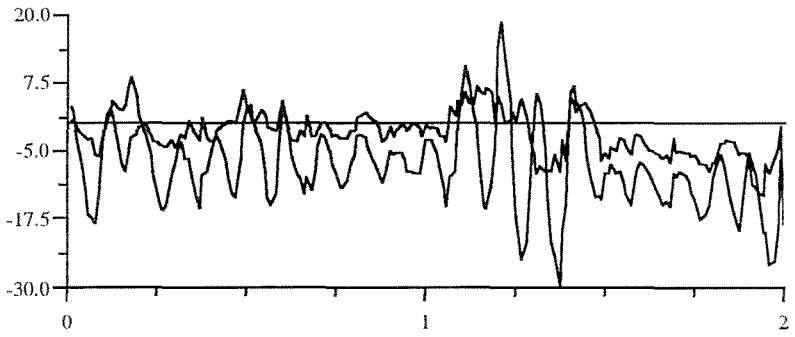

Figure 11: *Residues under misspecification and correct specification* $n = 1000$, $\theta_k = 1.0$.

A model of order 25 is tested first. The model barely identifies the underspecification. The modified Kolmogorov-Smirnov test is significant at the level 0.0445 (0.0086 in the case of $n = 5000$). The unmodified Kolmogorov-Smirnov test is significant at the level 0.413. Another simulation, using 0.5 instead of 1.0 as the coefficient of the extra term, could not identify at $n = 1000$.

The next figures show the residuals within the estimation interval. In figure 11 the lines of the misspecified and the sufficient specified models are drawn. The improvement it is obvious. From the 'misspecified' line the order of the missing

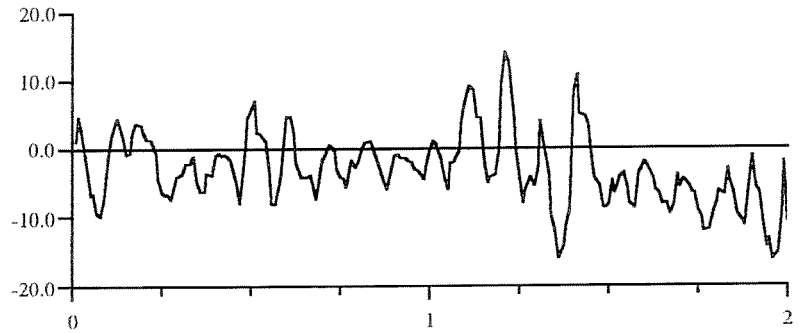

Figure 12: *Residues under misspecification* $n = 1000$, $\theta_k = 0.5$.

term can be counted. This will not be as easy in general presumably. In figure 12 the case with $\theta_k = 0.5$ is drawn. The residues are less in magnitude. They turn out to be insignificant (level 0.348) in the modified Kolmogorov-Smirnov test.

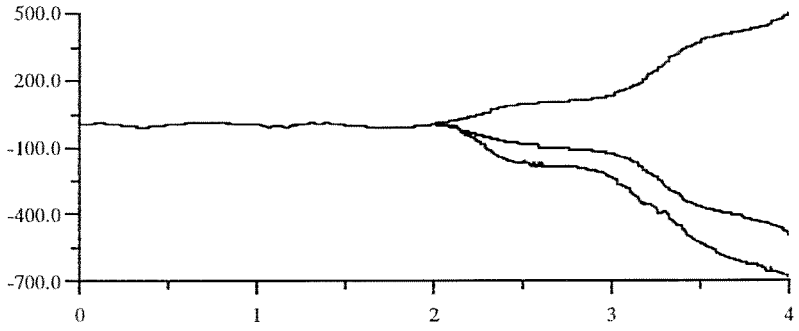

Figure 13: *Cumulative residual graph of* $n = 1000$, *underspecified case with* $\theta_k = 0.8$.
Cumulative residues indicate deviance of the model in the most obvious manner. The line on the bottom depicts the cumulative residues. Extending the model improves the results importantly.

Following these results an intermediate model is simulated using $\theta_k = 0.8$. This should be a barely identified case, which turned out to be the case. Then predictions are drawn from this model. It should be shown that missing out the term because it was not identified may not be that catastrophic. Two graphs are drawn, figure 13 containing the results from a $n = 1000$ (the actual number in $[0, 2]$ was $n = 402$) case with $\theta_k = 0.8$. The model is obviously insufficient for the period $[2, 4]$. The cumulative residues indicate this, again on the lower bound. Extending the model to satisfy the gradient test yields a sufficient model for $[2, 4]$. These results are graphed in figure 14. From figure 13 it is clear that in case predictions have to be made, estimating up to some time *before* the time point of which the last data are available, is very important.

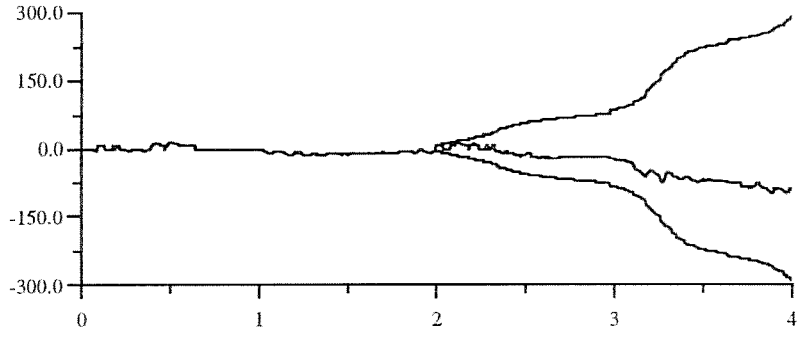

Figure 14: *Cumulative residual graph of* $n = 1000$, *underspecified case with* $\theta_k = 0.8$ *in extended setting.*
Cumulative residues indicate no deviance of the model. Extending the model improved the results considerably, compared to the not extended case. It must be noted that the margins are much wider than those generated by the not extended case.

Term	θ
Trend	-3.3427
Level	53.4323
$\sin(2\pi\, t)$	5.6356
$\cos(28\pi\, t)$	1.5421
$\sin(31\pi\, t)$	-2.0013

Table 3: *Parameters and terms used to test a non-exponential model*

5.4.3 *The quadratic case*

This simulation is used to see whether the model works acceptably if the true intensity function cannot be written as is supposed. Instead of a exponential function, a quadratic function is used, retaining the positiveness of the intensity function. As is done in the $n \to \infty$-case, the analysis is carried out in conjunction with a prediction. Again, points are sampled in the interval $[0, 4]$ and the models are estimated using the information in $[0, 2]$. Then the period $[2, 4]$ is predicted. It already turned out that this can be too big a period. In table 3 the parameters are listed.

The model actually fits the following:

See figure 15 for a comparison of the estimated model and the simulated model.

In figure 16 the cumulative residues are graphed. The residues break out at the end of the interval. See figure 17 for the integrated predictions.

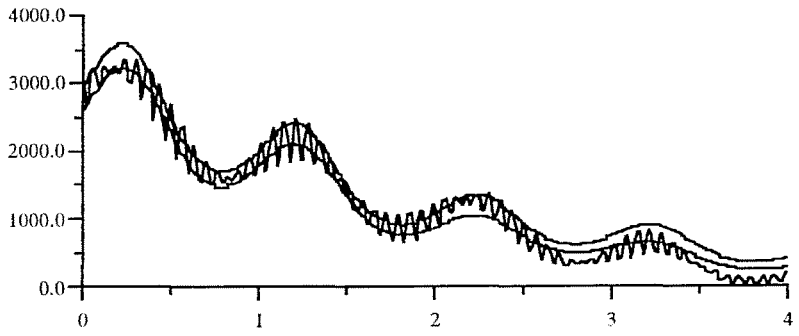

Figure 15: *Predictions of* $n = 5000$, *quadratic case.*
Predictions by the approximate model of table 4. Upper and lower bounds of the estimated exponential model are graphed together with the actual simulated quadratic model.

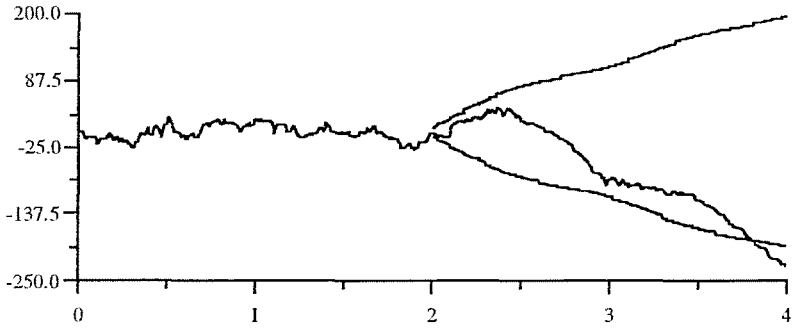

Figure 16: *Cumulative residual graph of* $n = 5000$, *quadratic case.*
Cumulative residues of the approximating exponential model of table 4. Only at the end of the prediction period the model beaks out the lowerbound. Another example of this feature. The extended model, convergence set at the gradient test as well does not solve this problem.

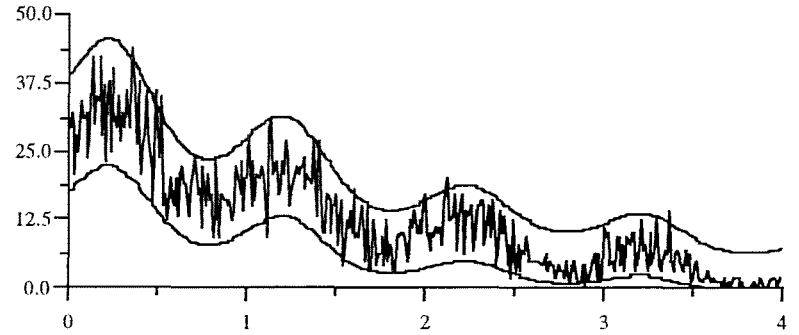

Figure 17: *Integrated Predictions of* $n = 5000$, *quadratic case.*
Integrated Predictions of the approximating exponential model. The observed counts rarely violate the confidence bounds in the estimation period. It can be doubted whether this model would have been very useful in practice except for the fact that the process is found to be periodic and decreasing. of table 4.

Term	θ
Const	8.007
Trend	-0.168
$\sin(2\pi\, t)$	0.293
$\cos(\pi\, t)$	-0.066

Table 4: *Parameters and terms estimated to fit a non-exponential model*

5.4.4 *Some conclusions*

In general the model seems to work well. In almost all cases its predictive power seemes to be sufficient, although in some examples the periodicity is not very well fitted. This particular problem exists mainly when few cycles have been observed. If the wave is substantial, but not too substantial, and the wavelength is long compared to the total period, sometimes the estimated wavelength is just off. This results in 'wave' residues. This means that only predictions based over a longer integrated period can have satisfactory precision.

Another observation is that the model in the limited number of observations case insufficiently indicates model deviance whereas the extended case, needing satisfaction of both the goodness of fit test and the gradient test works better. It is probably caused by lack of power of the goodness of fit in the small sample case because this phenomenon does not surface in case of more observations. In case of similar simulations it seems not to happen anymore at about $n = 1000$ observations really in the estimation procedure. In general this will depend on both the number of terms and their values (and their interrelations).

6. Real life example

The maximum likelihood model is not designed to estimate simulated data. In this section real data, consisting of injury accidents, are analysed. One problem that shows up is the fact that the information necessary for an analysis is not available in the distant past. Presently, information is available since 1979 and in some respect, since 1976. This means theoretically data from 1976 to the first half of 1993 can be analysed. In practice a shorter period of data has to be analysed. This is because consistency in the data is needed, interpretations change over time, and the registration lags for some accidents. In recent times, a major upgrade in the data structure took place by 1-1-1983, which is now chosen as t_0, the starting point. This recent starting point precludes analysis of safety-belt data and alcohol legislation.

A second consideration is that no exogenous variables can yet be included. This means that any planned analysis of an intervention should involve an intervention that did not influence the mobility of any kind or the use of particular modes of transportation. This restriction implies analysis of interventions that are not surrounded by important changes in usage of traffic. This happened for instance in the oil-crisis in the early seventies, which is too early anyway, or a quite recent possible intervention is the so called 'ov-jaarkaart' (season-ticket valid for all public transport) for students. This seemed, not thoroughly proven yet, to have reduced use of mopeds significantly, in favor of public transport. Preliminary analysis indicates a sharp decrease in the total number of casualties since introduction. But this happens in the younger age group as well. There could be some other explanation too. This example cannot be analysed until this lack of information is resolved. The German data in figure 1 are unavailable to the author at this level of disaggregation. Otherwise this would have been an ideal option.

An option available is the introduction of reflective bands on or in wheels of bicycles. Use of these became mandatory at 1-1-1987, but introduction progressed slowly. It was anticipated that the measure would not influence the use of the bicycles at all and would only have an influence on side accidents in which the bicycles are hit from the side. It is also assumed that the main effect of the measure is in twilight or darkness, which is the most dangerous period of the day for bicyclists.

The analysis of the data is used as an example, it is not intended as an traffic safety analysis, in which many more considerations have to be made.

The analysis is based on two groups, the side impact group and the non-side impact group. The definition of side impact is based on the manoeuvre, not based on the physical point of impact on the bicycle. This is registered, (as far as it is reliable) but it can be very misleading in this application. The analysis is based on the first 3 years, that is 1983, 1984 and 1985, the the lead period before the supposed intervention is 1986, just before the intervention took place in 1987. The first 3 year period produced 12277 side impact accidents, neglecting the few two-bicycle accidents in the dataset that are due to the poor lighting capacities probably not influenced by the intervention. The number of non-side

impact accidents was 26628 in the same period.

After some experimentation, a period of 7 years was found to be a reasonable base period. This period is translated in a time of 2π in the analyses. The terms listed in the respective tables reflect the fact. The side impact solution is listed in table 5. The non-side impact solution is listed in table table 6.

term	θ	term	θ
Const	2.387	$\sin(182t)$	4.660×10^{-2}
$\sin(7t)$	-1.256×10^{-1}	$\sin(141t)$	-4.832×10^{-2}
$\sin(28t)$	-9.626×10^{-2}	$\cos(52t)$	-4.918×10^{-2}
$\cos(64t)$	9.491×10^{-2}	$\sin(98t)$	4.700×10^{-2}
$\cos(7t)$	-9.209×10^{-2}	$\sin(23t)$	4.577×10^{-2}
$\cos(35t)$	-7.951×10^{-2}	$\sin(44t)$	4.580×10^{-2}
$\sin(48t)$	7.818×10^{-2}	$\cos(41t)$	-4.417×10^{-2}
$\cos(15t)$	1.090×10^{-1}	$\sin(80t)$	-4.252×10^{-2}
$\sin(9t)$	7.282×10^{-2}	$\sin(189t)$	4.207×10^{-2}
$\cos(14t)$	-7.288×10^{-2}	$\sin(130t)$	4.184×10^{-2}
$\cos(112t)$	-6.927×10^{-2}	$\cos(12t)$	4.367×10^{-2}
$\sin(66t)$	-6.796×10^{-2}	$\sin(127t)$	-4.073×10^{-2}
$\cos(105t)$	-6.263×10^{-2}	$\cos(57t)$	3.883×10^{-2}
$\sin(14t)$	-7.671×10^{-2}	$\sin(105t)$	3.840×10^{-2}
$\cos(115t)$	5.157×10^{-2}	$\cos(2t)$	4.055×10^{-2}
$\sin(146t)$	-4.932×10^{-2}		

Table 5: *The solution of the model based on the non-side impact accidents(3j/7j)*
The period of 7 years is translated in a time of 2π. All terms are listed based on the Wald test. The terms on top are tested most significant. All terms tested significant. The adapted Kolmogorov-Smirnov test was (barely) significant at the 0.050265049 level. Also a large number of points seem to be necessary. The gradient test was not satisfied. A cumulative residual plot can be found in figure 18.

All terms listed tested significant on the Wald test. The terms on top tested most significant. The adapted Kolmogorov-Smirnov test was (barely) significant at the 0.050265049 level in the side impact case. Also a large number of terms seemed to be necessary to get a sufficient fit. A cumulative residue plot can be found in figure 18.

The non side impact case yielded a more attractive solution. The solution tested well on the adapted Kolmogorov-Smirnov test, the test was significant at the 0.117284009 level, and needed not much terms. A cumulative residue plot can be found in figure 19.

The gradient test was not satisfied in either model.

What was hoped to see in figure 18 can be seen in figure 19, depicting the number of side accidents with at least one bicyclist involved. Unfortunately, even

term	θ	term	θ
Const	3.128	$\cos(1t)$	5.755×10^{-2}
$\cos(7t)$	-3.143×10^{-1}	$\cos(24t)$	-1.051×10^{-1}
$\sin(7t)$	-1.903×10^{-1}	$\sin(28t)$	-7.101×10^{-2}
$\sin(48t)$	1.056×10^{-1}	$\sin(63t)$	-6.427×10^{-2}
$\cos(15t)$	1.295×10^{-1}	$\cos(76t)$	3.898×10^{-2}
$\sin(34t)$	7.839×10^{-2}	$\sin(25t)$	9.501×10^{-2}
$\cos(14t)$	-7.956×10^{-2}	$\sin(14t)$	-4.284×10^{-2}
$\cos(141t)$	-5.937×10^{-2}	$\cos(29t)$	4.327×10^{-2}
$\cos(26t)$	1.079×10^{-1}	$\cos(64t)$	2.756×10^{-2}

Table 6: *The solution of the model based on the non-side impact accidents(3j/7j)*
The period of 7 years is translated in a time of 2π. All terms are listed based on the Wald test. The terms on top are tested most significant. All terms tested significant. The adapted Kolmogorov-Smirnov test was significant at the 0.117284009 level. The gradient test was not satisfied. A cumulative residual plot can be found in figure 19.

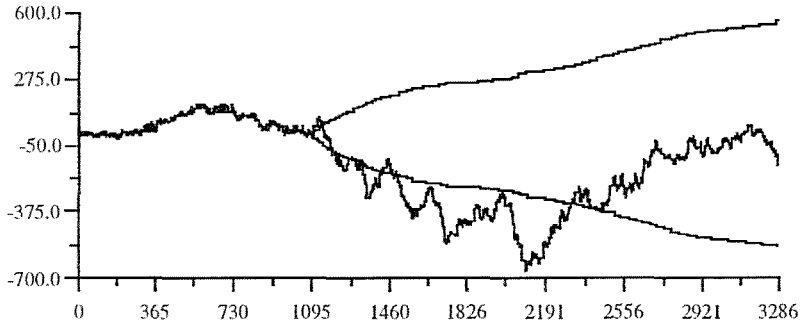

Figure 18: *Residues of bicyclist side-impact*
This model was estimated based on the first three years, 1983 – 1985. Around day 1460 the intervention should start its influence. Unfortunately, this is not clearly visible, if at all. The wave in the estimated period may indicate a basic wave not in the model. Inspecting the whole period, probably a 14-year base period may be needed. It has to be doubted however, in recollection of the results in the simulation study, that such a wave can be estimated reliably from the data.

after specific checking, this graph depicts the *non*-side impact accidents. The effect of intervention should have started at around day 1462 (1987), so the first year after the estimation period seems to be predicted well by the model. Somewhere in the second halve of 1987 the model seems to divert slightly. This may be caused by the short (3-year) estimation period. The sharp decrease in the summer of 1988 may be the result of something special. It is thought that this cannot be caused by the deteriorating prediction quality, because of the sudden change. There seems to be no explanation of this phenomenon. The sudden decrease can also be found in the side-impact accident residues. In this case the peak is rather obscured by other peaks.

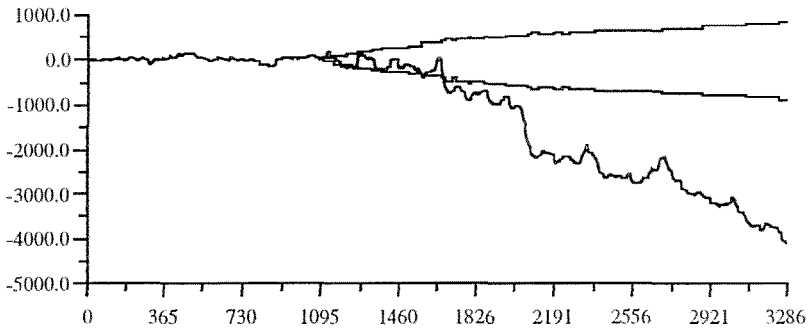

Figure 19: *Residues of bicyclist non side impact*
What can be seen here is what was hoped to be seen in figure 18, depicting the number of side accidents with at least one bicyclist involved. Unfortunately, even after specific checking, this graph depicts the *non*-side impact accidents. The effect of intervention should have started at around day 1462, so the first year after the estimation period seems to be predicted well by the model. Somewhere in the second halve of 1987 the model seems to divert slightly. This may be caused by the short (3-year) estimation period. The sharp decrease in the summer of 1988 may be the result of something special. It is thought that this cannot be caused by the deteriorating prediction quality, because of the sudden change. Because the residues seem to proceed as before after some time, it seems there has been a short period in which a relatively small number of accidents occurred. This effect can also be seen in the side impact accidents, although it is rather obscured by other disturbances.

From the traffic safety point of view, no conclusion can be drawn from these results. This unfortunate result is followed by the conclusion that problems may have been caused by specification inflexibility. The problem in figure 18 may have been caused by the fact that the system of functions is too tight. It seems advisable to allow the user to specify certain functions ahead. For instance it could be useful if the user can apply multiple systems to the model such as the base model plus, in this case, a number of terms with wavelength of about 14-years. This would limit the total number of estimable parameters. There is no reason not to do this unless the total number of terms is too much and one needs the scheme in § 3.7.4.

7. Conclusions

From the results so far it can be concluded that the project is not completed. The model-method combination suffers from some shortcomings and could be improved at a number of points.

The most eye-catching improvement would be improving the goodness of fit test. It has already been seen that it does not seem to be very powerful in the current implementation. Confidence in the model would improve if the power of the test were increased. Some directions of research are available based on Durbin [1973] and Pollard [1984].

Another point may be found in the general application of (asymptotical) normality assumptions. In § 4.2.2 a tendency toward 'breaking out' through the lower limit of the cumulative residuals seems to show up. This may be an indication of a non-symmetrical distribution of those residues. Normality is assumed in § 4.2.2. This phenomenon should be studied further. Although not observed, this phenomenon will occur in other situations as well. In general, it could be studied if small sample tests can be derived for some of the tests used here. In practice this may not be possible in all cases, if useful at all.

Not many robustness considerations have been applied so far. It has been found that a misidentification of a seasonal effect had serious consequences on the long-term predictions. Apart from the question of whether to use such predictions at all, it can be argued that small-term deviations in the data can have a long term effect. This is partly due to the Fourier system in use and essentially a consequence of long range functions.

The total number of parameters to be estimated may cause a statistical problem on it's own. It may be argued that this implementation is essentially a infinite number of parameters problem, because no strict maximum of the complexity is used or that maximum is 'estimated'. It should be considered whether or not this invalidates some of the assumptions.

Finally, by defining a good general goodness-of-fit criterion, this criterion can be used to validate the original count data techniques as well. The goodness-of-fit criterion itself thus may be very useful in practice, with or without a system of functions.

In the conclusion, in § 3.7.5 (Selecting more terms in one time) the use of the lagrange-multiplier test is omitted. It could be useful to test a increasingly longer version of $\hat{\theta}$ until either the lagrange multiplier test indicates insignificance or the maximum number of terms that are allowed to be included is reached.

Bibliography

Agostino, R.B.D. & Stephens, M.A. (1986). *Goodness-of-fit tests*. Number 68 in Statistics, textbooks and monographs. Marcel Dekker, New York.

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. In Petrov, B.N. & Csaki, F., editors, *Proceedings of the Second International Symposium of Information of Information Theory*, pages 267–281, Budapest. Academica Kiado, Academica Kiado.

Box, G.E.P. & Tiao, G.C. (1975). *Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association* 70(349), 70–79.

Braun, H. (1980). *A simple method for testing goodness of fit in the presence of nuisance parameters. Journal of the Royal Statistical Society* 42, 53–63.

Cheng, R. & Stephens, M.A. (1989). *A goodness-of-fit test using Moran's statistic with estimated parameters. Biometrika* 76(2), 385–92.

Durbin, J. (1973). *Distribution Theory for Tests Based on the Sample Distribution Function*, volume 9 of *Regional conference series in appl. math.* Siam, Philadelphia.

Epps, T.W. & Pulley, L.B. (1983). *A test for normality based on the empirical characteristic function. Biometrika* 70(3), 723–728.

Ernst, G. & Brüning, E. (1990). *Fünf Jahre danach:, Wirksamkeit der 'Gurtanlegepflicht für Pkw Insassen ab 1. 8. 1984'. Zeitschrift für Verkehrssicherheit* 36(1), 2–13.

Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and survival analysis*. John Wiley & Sons, New York.

Fletcher, R. (1981). *Practical methods of optimization*, volume II. John Wiley & Sons, Chicester.

Good, I.J. & Gaskins, R. (1971). *Nonparametric roughness penalties for probability densities. Biometrika* 58(2), 255–277.

Grandell, J. (1990). *Aspects of risk theory*. Springer-Verlag, Berlin, Heidelberg.

Grenander, U. (1981). *Abstract Inference*. John Wiley & Sons, New York. Method of sieves.

Hampel, F.R.; Rousseeuw, P.J.; Ronchetti, E.M., & Stahel, W.A. (1986). *Robust statistics*. John Wiley & Sons, New York.

Härdle, W. (1990). *Smoothing techniques*. Springer-Verlag, Berlin, Heidelberg.

Harvey, A.C. (1985). *Multivariate time series models, control groups and intervention analysis*. Economics Programme Discussion Paper A53, London School of Economics, London.

Harvey, A.C. (1986). *The effects of seat belt legislation on British road casualties: A case study in structural time series modeling. Journal of the Royal Statistical Society* 149, 187–227.

Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical analysis of failure time data*. John Wiley & Sons, New York.

Karr, A.F. (1991). *Point Processes and their statistical inference*. Marcel Dekker, New York, second edition.

Kendall, M. & Ord, J.K. (1990). *Time series*. Edward Arnold, London, third edition.

Kendall, M. & Stuart, A. (1987). *Advanced Theory of Statistics*, volume 2. Charles Griffin & Co, London.

Kullback, S. (1968). *Information Theory and Statistics.* John Wiley & Sons, New York.

Luenberger, D.G. (1984). *Linear and nonlinear programming.* Addison-Wesley, Reading, Massachusetts, second edition.

Moran, P. (1951). *The random devision of an interval-part ii. Journal of the Royal Statistical Society B* 13, 147–150.

Pollard, D. (1984). *Convergence of stochastic Processes.* Springer series in statistics. Springer-Verlag, New York.

Press, W.H.; Flannery, B.P.; Teukalsky, S.A., & Vetterling, W.T. (1989). *Numerical recipes in Pascal.* Cambridge University Press, Cambridge.

Rao, K.C. (1972). *The Kolmogorov-Smirnov, Cramer-von Mises, chisquare statistics for goodness-of-fit in the parametric case. Bull. Inst. Math. Statist.* 1, 67. Abstract 133-6.

Serfling, R.J. (1980). *Approximation theorems of mathematical statistics.* Wiley series in probability and mathematical statistics. John Wiley & Sons, New York.

Snyder, D.L. & Miller, M.I. (1990). *Random Point Processes in time and space.* Springer-Verlag, New York.

Statistisches Bundesamt. (1988). *Verkehr, Verkehrsunfälle.* Fachserie 8, reihe 7, Statistisches Bundesamt, Wiesbaden.

Stephens, M.A. (1986). *Tests based on EDF statistics*, chapter 4, pages 97–194. In Agostino & Stephens [1986].

Stoer, J. & Bulirsch, R. (1980). *Intoduction to numerical analysis.* Springer-Verlag, New York.

Wald, A. (1943). *Test of statistical hypothesis concerning several parameters when the number of observations is large. Trans. Am. Math. Soc.* 54, 426–482.

White, H. (1982). *Maximum likelihood estimation of misspecified models. Econometrika* 50(1), 1–25.

Appendix A. Figures

A.1 $n \to \infty$ simulation

Following are graphs of the predictions of the maximum likelihood model. All graphs contain the upper and lower limits based on the estimated model, together with the simulated intensity functions. From this view, it seems the model estimated the intensity well.

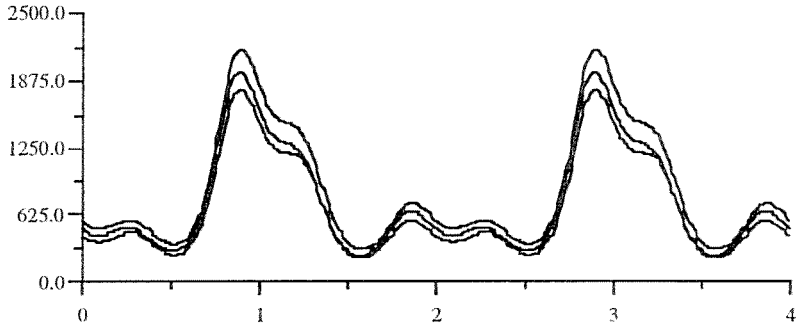

Figure 20: *Predictions and true intensity of* $n = 3000$.

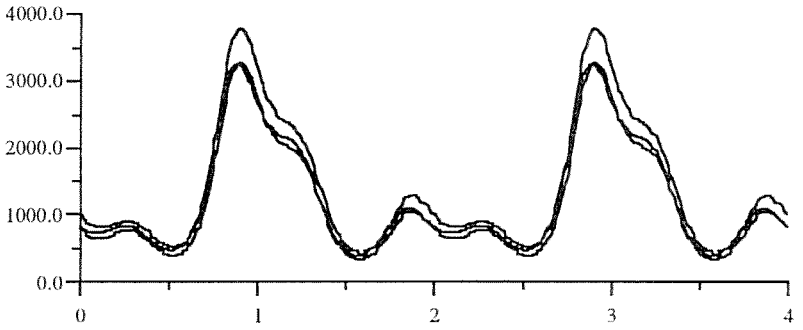

Figure 21: *Predictions and true intensity of* $n = 5000$.

Following are the Cumulative residuals of the $n \to \infty$ simulation. The line over the entire length is the cumulative residual. The lines starting off 2 are its upper confidence level and the lower confidence level.

The last feature graphs of the case $n \to \infty$ contain the results of the integrated intensity functions. The interval length is 0.01, thus delivering 400 intervals. For every interval the number of simulated points are counted. Also both the upper and lower 95% confidence limits based on the respective models for these intervals are computed. Only in rare cases the number of points seem out of the confidence bounds. This number does not question the validity of the confidence limits. It seems the region could be smaller still.

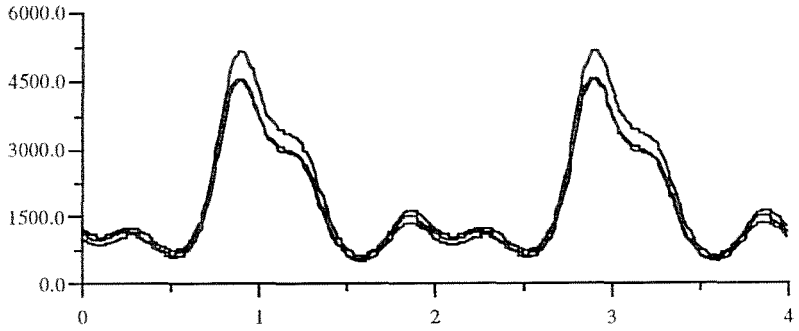

Figure 22: *Predictions and true intensity of* $n = 7000$.

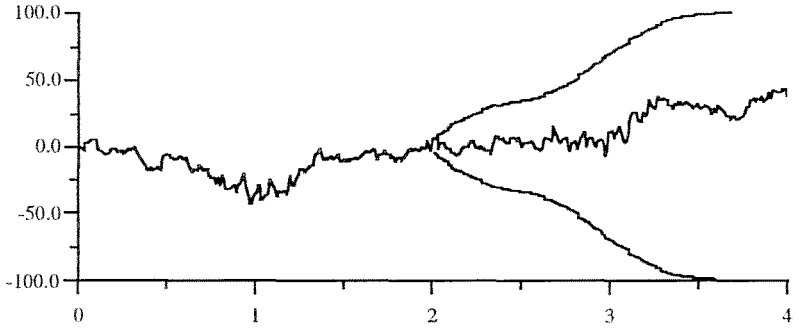

Figure 23: *Cumulative residual graph of* $n = 3000$.

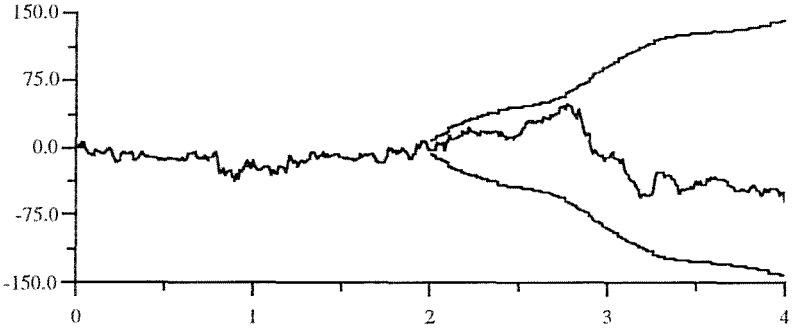

Figure 24: *Cumulative residual graph of* $n = 5000$.

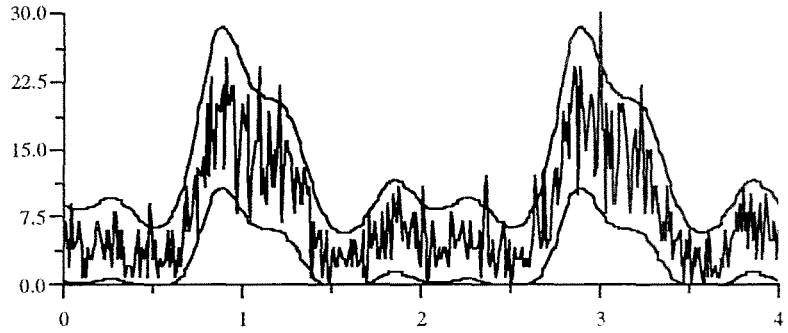

Figure 25: *Graph of integrated prediction interval and tabulated points* $n = 3000$.

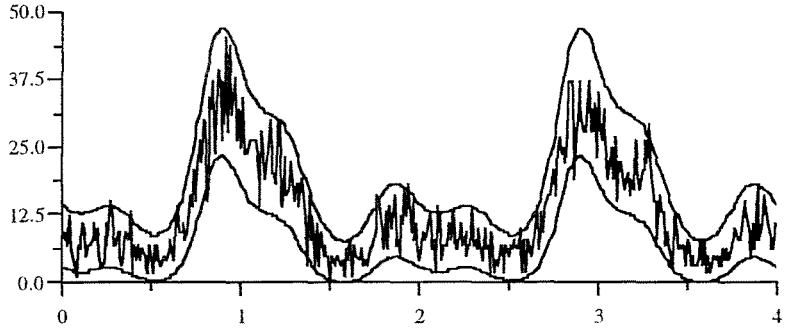

Figure 26: *Graph of integrated prediction interval and tabulated points* $n = 5000$.