Odhad spolehlivosti kolejových obvodů z nekompletních cenzorovaných dat

Petr Novák, Rudolf Blažek, Martin Daňhel

Katedra aplikované matematiky, Fakulta informačních technologií ČVUT

23. ledna 2018

Úvod

- Na traťových úsecích v ČR je přítomnost vlaků detekována pomocí kolejových obvodů.
- Kolejové obvody jsou složeny z několika komponent.
- Chceme odhadnout rozdělení doby do poruchy komponent na základě data o poruchách z provozu z let 2006-2016.

Detekce vlaků

Bodové detektory:

- zátěžové – počitadlo os,
- elektronické,
- rádiové – Eurobalise, viz dále.

Detekce vlaků

Bodové detektory:

- zátěžové – počitadlo os,
- elektronické,
- rádiové – Eurobalise, viz dále.

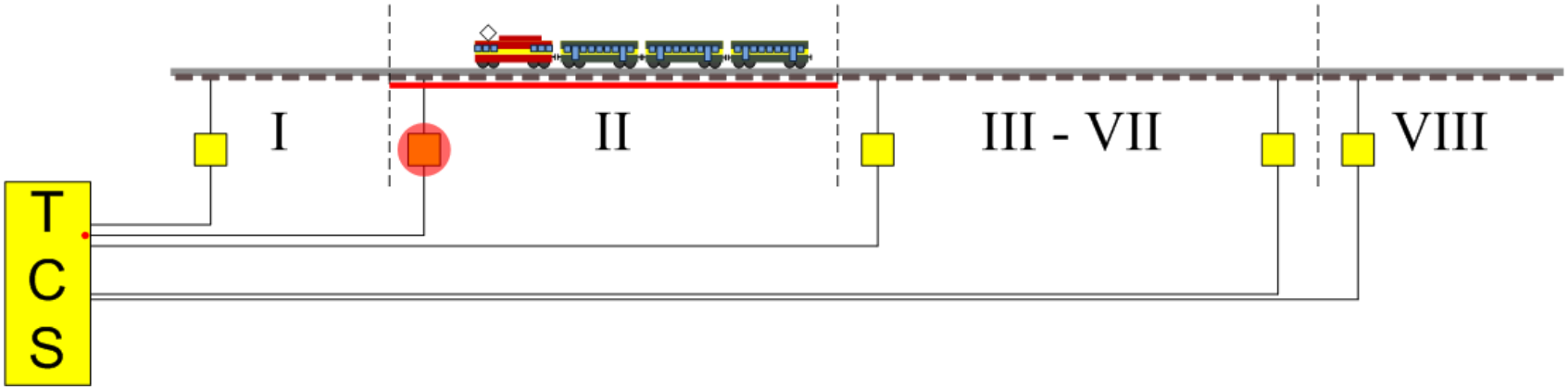
Úseková detekce – kolejové obvody

- Elektrický proud skrz kolejnice.
- V ČR - kolejový obvod KOA-1, výrobce AŽD Praha.
- Elektronické vyhodnocení.
- Na mnoha úsecích železnice i v metru.
- Vysoké nároky na spolehlivost.

Kolejové obvody

Kolejový obvod (track circuit):

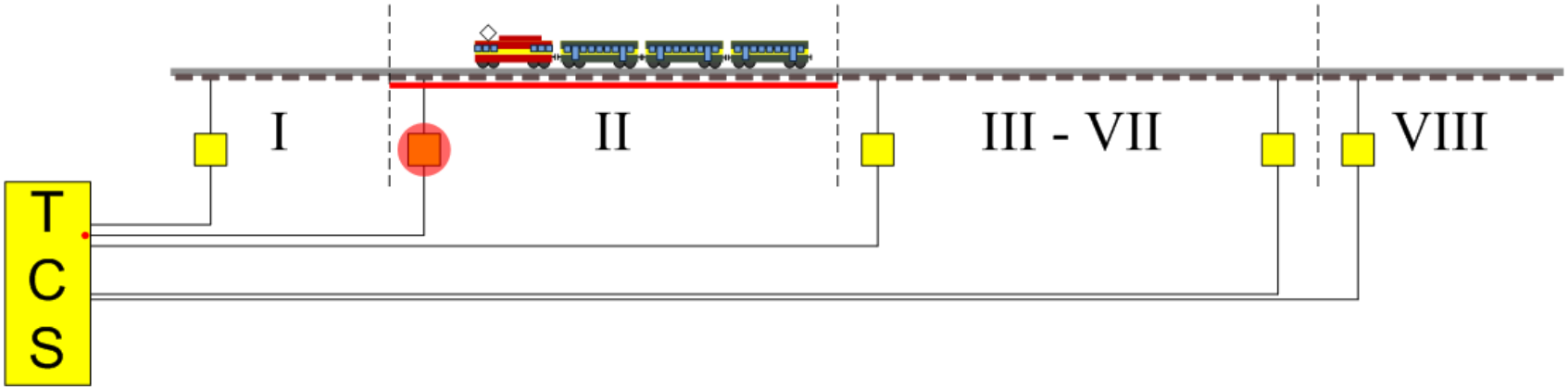
- Detekční část
- Vyhodnocovací část - jedna stanice pro až osm bloků



Kolejové obvody

Kolejový obvod (track circuit):

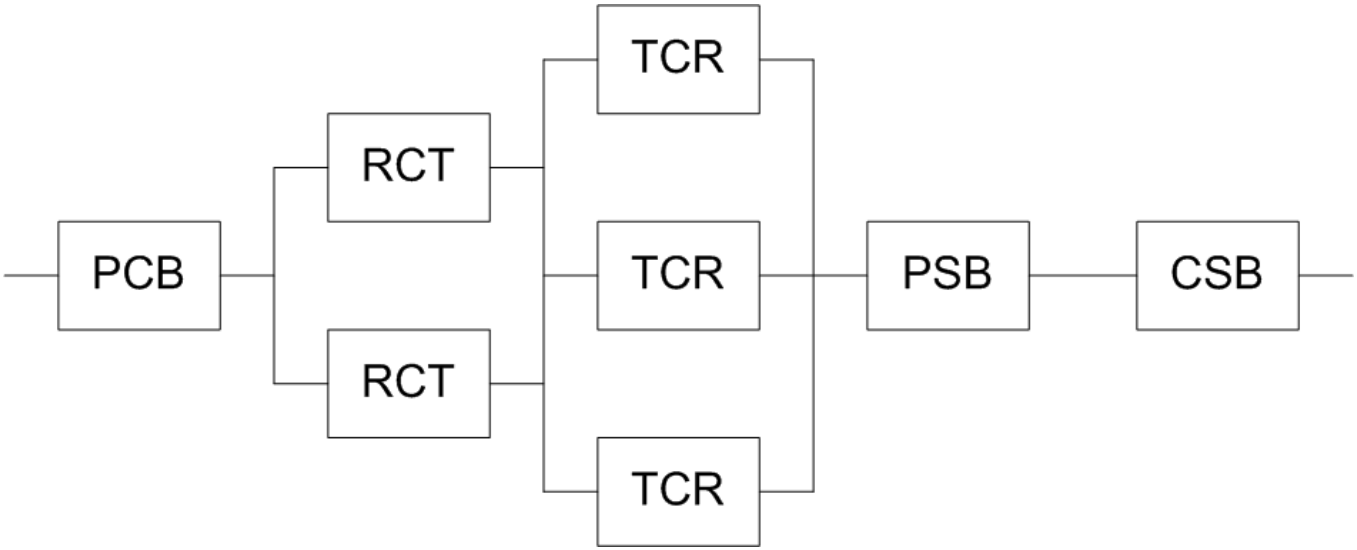
- Detekční část
- Vyhodnocovací část - jedna stanice pro až osm bloků



Bezpečnostní požadavky

- Kolejové obvody musí splňovat bezpečnostní požadavky SIL-4 (Safety Integrity Level) dle normy EN 50126.
- Výměna vadných komponent do 80 hodin.
- Požadovaná střední životnost 20 let.

Vyhodnocovací část kolejového obvodu



PCB – Jistič
RCT – Usměrňovač – redunantní, k provozu stačí jeden
TCR – Track Circuit Receiver – redunantní, k provozu stačí dva
PSB – Rozvodná deska
CSB – Komunikační deska

Popis dat

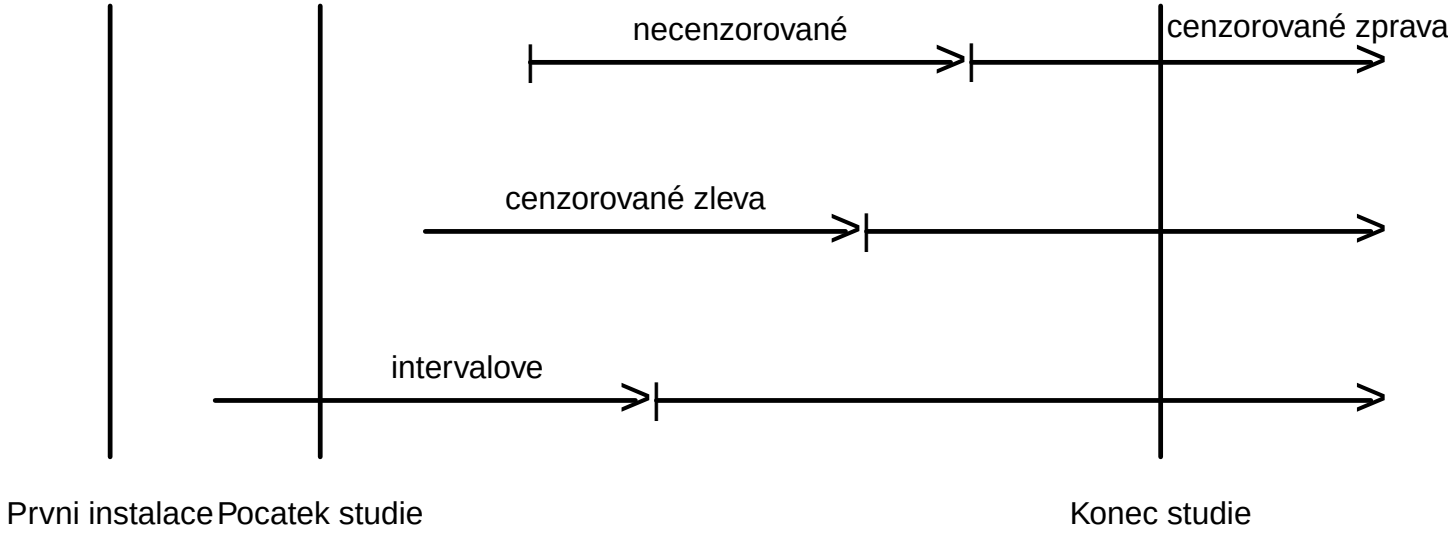
Servisní databáze

- Záznamy výrobce o poruchách a opravách
- Záznamy z 2006 až 2016
- Počet instalovaných zařízení narůstal
- V případě poruchy byla komponenta vyměněna za novou, v případě softwarové chyby byl proveden rework.
- Po opravě je součástka vždy uvažována jako nová
- Nejdůležitější část – TCR, provádí samotné vyhodnocení.

Chceme modelovat rozdělení doby do poruchy.

Cenzorovaná data

- U některých stanic chyběl údaj o uvedení do provozu. Bylo možné jen identifikovat, kdy ještě stanice nebyla funkční $\Rightarrow$ cenzorování zleva.
- Někdy bylo možné zjistit možný interval uvedení do provozu $\Rightarrow$ intervalové cenzorování.
- Data jsou k dispozici jen do konce roku 2016 $\Rightarrow$ cenzorování zprava.



Nekompletní data

- Některé poruchy byly přiřazeny k traťovému úseku a ne ke stanici
- Nebylo rozlišováno mezi různými zařízeními v rámci jedné stanice.
- Potřeba ošetřit.
- Kdybychom znali rozprostření poruch mezi stanice a zařízení, mohli bychom použít běžné nástroje analýzy přežití.
- Parametricky – metoda maximální věrohodnosti.
- Neparametricky – Kaplan-Meierův odhad.

Parametrické modelování doby do poruchy

Možnosti vyjádření rozdělení:

- Funkce přežití $S(t) = \mathrm{P}(T > t)$.
- Intenzita poruch $\lambda(t) = \lim_{h \to 0} P(t \leq T \leq t + h | T \geq t)/h$.
- Vztah: $\lambda(t) = f(t)/S(t)$.

Parametrické modelování doby do poruchy

Možnosti vyjádření rozdělení:

- Funkce přežití $S(t) = \mathrm{P}(T > t)$.
- Intenzita poruch $\lambda(t) = \lim_{h \to 0} P(t \leq T \leq t + h | T \geq t)/h$.
- Vztah: $\lambda(t) = f(t)/S(t)$.

Odhad rozdělení:

- Parametricky – maximalizováním věrohodnosti. Příspěvek jednoho pozorování je

$$g(T) = f(T)^{\Delta_0} F(T - T_0)^{\Delta_L} \left(F(T - T_0) - F(T - T_1)\right)^{\Delta_I} S(T)^{\Delta_R}.$$

- Δ_0 – indikátor necenzorovaného selhání,
- Δ_L – indikátor cenzorování zleva,
- Δ_I – indikátor intervalového cenzorování,
- Δ_R – indikátor cenzorování zprava.

Parametrické odhady

Postup:

- Prozkoumáme pro několik parametrických rodin:
  - exponenciální s $\lambda(t) = \lambda$,
  - Weibullovo s $S(t) = e^{-(\lambda t)^a}$,
  - gamma s $f(t) \propto t^{a-1} e^{-\lambda t}$,
  - lognormální.
- Odhadneme parametry pomocí metody maximální věrohodnosti.
- Podle hodnoty věrohodnosti můžeme porovnat, které rozdělení popisuje data nejlépe.

Parametrické odhady – věrohodnost

Uvažujme:

- N stanic
- V každé stanici k_s zařízení.
- Z každé stanice záznam o n_s poruchách v časech $T_{s1}, \ldots, T_{sn_s}$.
- Rozprostření poruch mezi zařízení pomocí indexů

$$\mathbf{z}_s = (z_{s1}, \ldots, z_{sn_s}).$$

Parametrické odhady – věrohodnost

Uvažujme:

- N stanic
- V každé stanici k_s zařízení.
- Z každé stanice záznam o n_s poruchách v časech $T_{s1}, \ldots, T_{sn_s}$.
- Rozprostření poruch mezi zařízení pomocí indexů

$$\mathbf{z}_s = (z_{s1}, \ldots, z_{sn_s}).$$

Celkem $k_s^{n_s}$ možností.

Parametrické odhady – věrohodnost

Uvažujme:

- N stanic
- V každé stanici k_s zařízení.
- Z každé stanice záznam o n_s poruchách v časech $T_{s1}, \ldots, T_{sn_s}$.
- Rozprostření poruch mezi zařízení pomocí indexů

$$\mathbf{z}_s = (z_{s1}, \ldots, z_{sn_s}).$$

Celkem $k_s^{n_s}$ možností.

Kdybychom znali rozprostření poruch, mohli bychom sestavit věrohodnostní funkci:

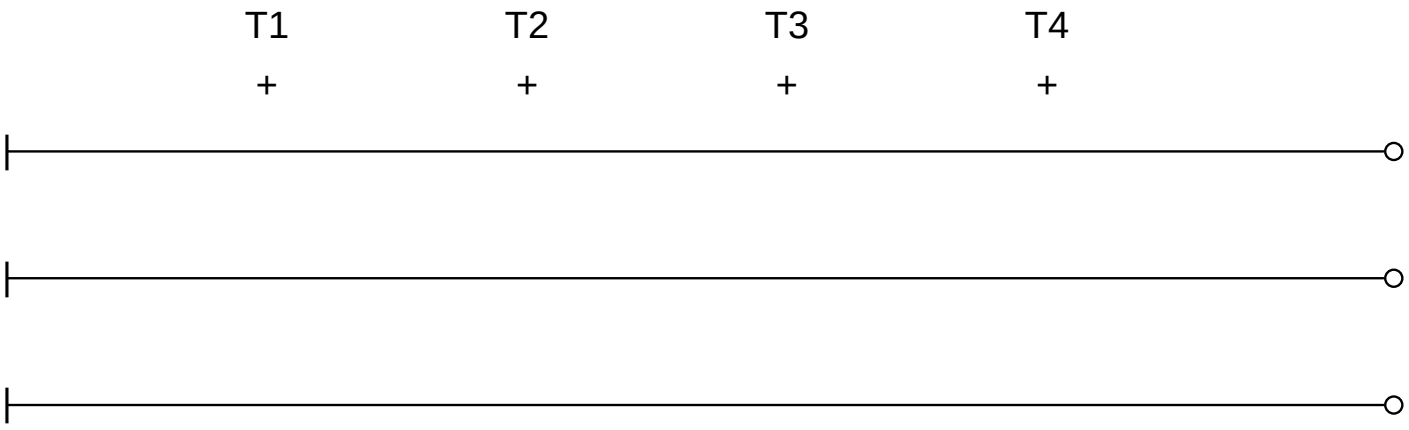
$$L(\theta, T_{\bullet\bullet}|\mathbf{z}_\bullet) = \prod_{s=1}^{N} \prod_{j=1}^{k_s} \prod_{i=1}^{n_s} g^{I(z_{si}=j)}(T_{si} - \max_{l=1,\ldots i-1}(T_{sl}, z_{sl} = j)).$$

Pro parametrický model pak můžeme rovnou maximalizovat.

Parametrické odhady – rozprostření poruch

Nevíme, která z poruch se týkala kterého ze zařízení. Možnosti:

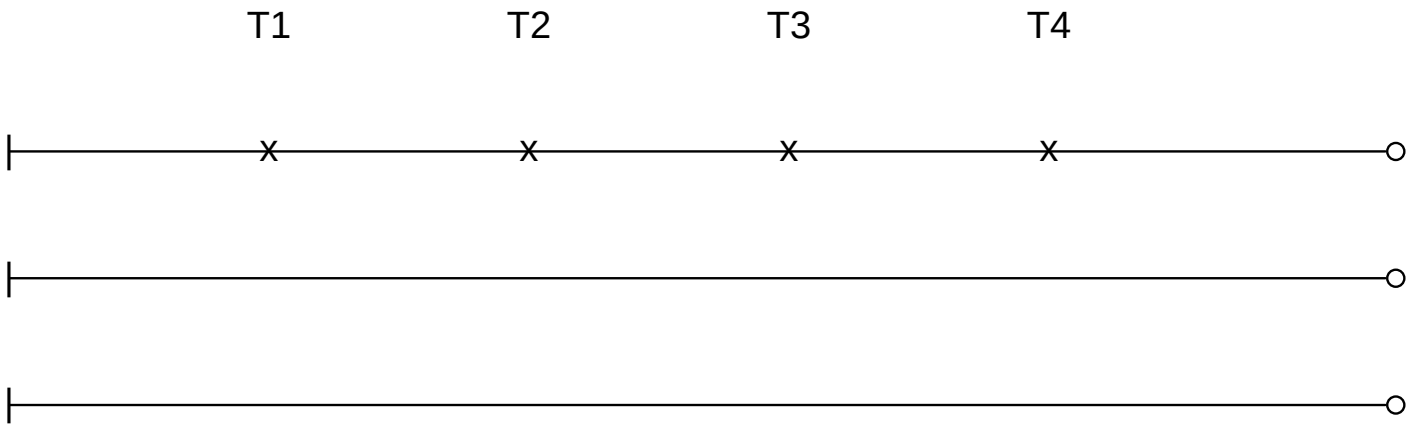
- Všechny na jednom zařízení
- Na všech rovnoměrně
- Náhodně.



Parametrické odhady – rozprostření poruch

Nevíme, která z poruch se týkala kterého ze zařízení. Možnosti:

- Všechny na jednom zařízení
- Na všech rovnoměrně
- Náhodně.

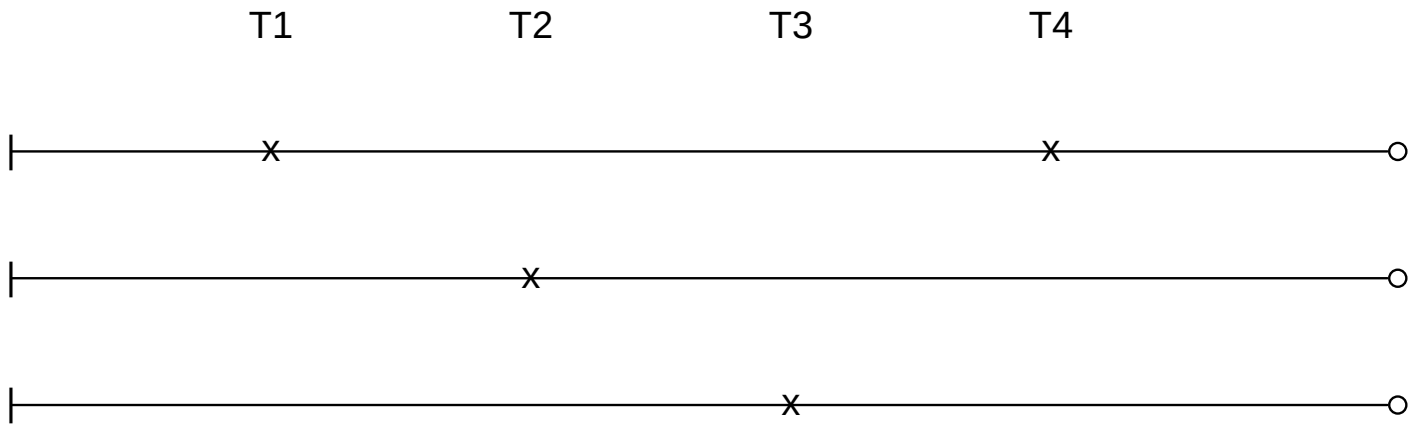


- Pro exponenciální rozělení vyjde stejně díky bezpaměťovosti.

Parametrické odhady – rozprostření poruch

Nevíme, která z poruch se týkala kterého ze zařízení. Možnosti:

- Všechny na jednom zařízení
- Na všech rovnoměrně
- Náhodně.



- Pro exponenciální rozělení vyjde stejně díky bezpaměťovosti.

Parametrické odhady – rozprostření poruch

Nevíme, která z poruch se týkala kterého ze zařízení. Možnosti:

- Všechny na jednom zařízení
- Na všech rovnoměrně
- Náhodně.

- Pro exponenciální rozělení vyjde stejně díky bezpaměťovosti.

Rozprostření poruch – možná řešení

Náhodně:

- Musíme předpokládat určité chování poruch, např. nezávislé rovnoměrné rozprostření.
- Vygenerujeme rozprostření a odhadneme parametry.
- Můžeme opakovat mnohokrát a odhady zprůměrovat.

Rozprostření poruch – možná řešení

Náhodně:

- Musíme předpokládat určité chování poruch, např. nezávislé rovnoměrné rozprostření.
- Vygenerujeme rozprostření a odhadneme parametry.
- Můžeme opakovat mnohokrát a odhady zprůměrovat.

Bayesovsky – EM algoritmus

- Věrohodnost uvažujeme jako směs přes všechna možná rozprostření.
- Iterativně střídavě odhadujeme pravděpodobnosti jednotlivých rozprostření a maximalizujeme věrohodnost.
- Analyticky i výpočetně náročné.

Rozprostření poruch – možná řešení

Náhodně:

- Musíme předpokládat určité chování poruch, např. nezávislé rovnoměrné rozprostření.
- Vygenerujeme rozprostření a odhadneme parametry.
- Můžeme opakovat mnohokrát a odhady zprůměrovat.

Bayesovsky – EM algoritmus

- Věrohodnost uvažujeme jako směs přes všechna možná rozprostření.
- Iterativně střídavě odhadujeme pravděpodobnosti jednotlivých rozprostření a maximalizujeme věrohodnost.
- Analyticky i výpočetně náročné.

Bayesovsky – MH algoritmus (MCMC)

- Určíme apriorní rozdělení parametrů a rozprostření poruch.
- Generujeme návrhy, které buďto přijmeme nebo zamítneme.
- Citlivé na volbu apriorního rozdělení.

Parametrické odhady – bayesovský přístup

- Apriorní rozprostření poruch mezi zařízení a parametry rozdělení považujeme za náhodné s hustotou $p(\theta, \mathbf{z})$.
- Rozdělení dat při pevných hodnotách parametrů $r(\mathbf{T}|\theta, \mathbf{z})$ odpovídá věrohodnostní funkci
- Aposteriorní rozdělení je pak

$$\pi(\theta, \mathbf{z}|\mathbf{T}) = \frac{r(\mathbf{T}|\theta, \mathbf{z})p(\theta, \mathbf{z})}{\int r(\mathbf{T}|\theta, \mathbf{z})p(\theta, \mathbf{z})}.$$

Parametrické odhady – bayesovský přístup

- Apriorní rozprostření poruch mezi zařízení a parametry rozdělení považujeme za náhodné s hustotou $p(\theta, \mathbf{z})$.
- Rozdělení dat při pevných hodnotách parametrů $r(\mathbf{T}|\theta, \mathbf{z})$ odpovídá věrohodnostní funkci
- Aposteriorní rozdělení je pak

$$\pi(\theta, \mathbf{z}|\mathbf{T}) = \frac{r(\mathbf{T}|\theta, \mathbf{z})p(\theta, \mathbf{z})}{\int r(\mathbf{T}|\theta, \mathbf{z})p(\theta, \mathbf{z})}.$$

Přímý výpočet složitý, použili jsme Metropolis-Hastingsův algoritmus:

- Uvažujeme neinformativní apriorní rozdělení rozprostření poruch.
- Vygenerujeme rozprostření poruch a návrhové parametry.
- Návrh přijememe s pravděpodobností rovnou poměru věrohodností pro nové a staré parametry.
- Opakujeme.

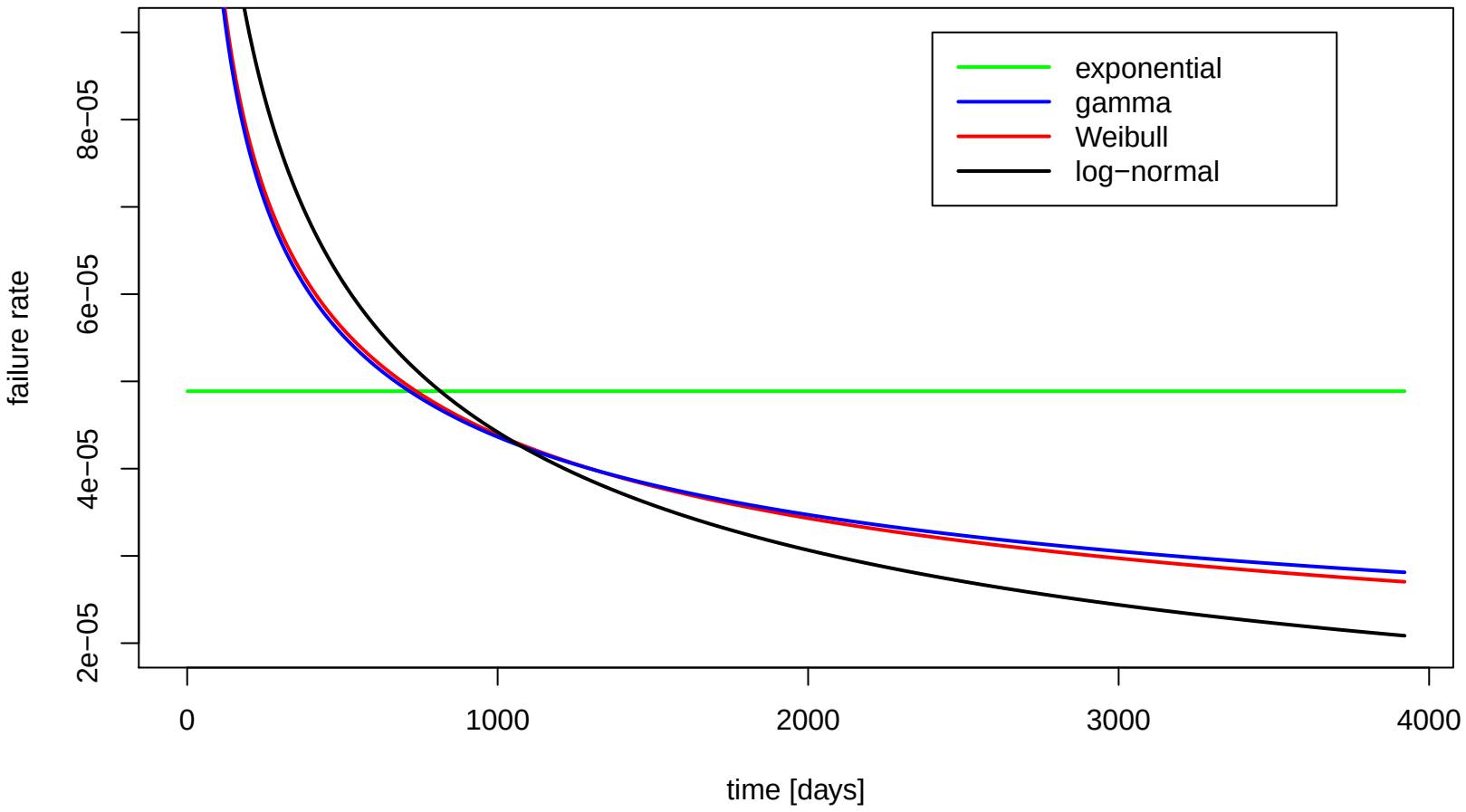
Odhadnuté parametry

- Odhadnuté parametry pro vyhodnocovací jednotku TCR.
- Data randomizována kvůli důvěrnosti.

algoritmus	rozd.	log-lik	$\hat{\lambda}$	$\hat{a}$
náhodně	ex	-2828	4.84e-05	–
	wb	-2787	1.25e-05	0.61
	ga	-2788	8.63e-06	0.59
	ln	-2783	$\hat{\mu} = 11.93$	$\hat{\sigma} = 3.42$
Metropolis- -Hastings	ex	-2830	5.44e-05	–
	wb	-2825	3.40e-05	0.75
	ga	-2827	3.35e-05	0.80
	ln	-2839	$\hat{\mu} = 11.33$	$\hat{\sigma} = 3.07$

- Lognormální rozdělení dává nejlepší výsledky co do věrohodnosti.
- M.-H. algoritmus dává o něco horší výsledky.
- Pro ostatní zařízení dává nejvyšší věrohodnost Weibullovo nebo lognormální.

Odhadnuté intenzity poruch jednotky TCR



Odhad životnosti

Můžeme odhadnout:

- Střední dobu do poruchy pro jednotlivá zařízení – dosazením odhadnutých parametrů.
- Střední dobu do poruchy celého systému.

Odhad životnosti

Můžeme odhadnout:

- Střední dobu do poruchy pro jednotlivá zařízení – dosazením odhadnutých parametrů.
- Střední dobu do poruchy celého systému.

Díky redundanci usměrňovačů (stačí jeden ze dvou) a vyhodnocovacích jednotek TCR (dva ze tří) můžeme dobu do poruchy vyjádřit jako

$$T_{\mathrm{CELK}} = \min(T_{\mathrm{PCB}}, T_{\mathrm{RCT(2:2)}}, T_{\mathrm{TCR(2:3)}}, T_{\mathrm{PSB}}, T_{\mathrm{CSB}}).$$

Považujeme-li součástky za nezávislé, funkci přežití minima lze přepsat jako součin funkcí přežití jednotlivých komponent.

Odhad životnosti

Můžeme odhadnout:

- Střední dobu do poruchy pro jednotlivá zařízení – dosazením odhadnutých parametrů.
- Střední dobu do poruchy celého systému.

Díky redundanci usměrňovačů (stačí jeden ze dvou) a vyhodnocovacích jednotek TCR (dva ze tří) můžeme dobu do poruchy vyjádřit jako

$$T_{\mathrm{CELK}} = \min(T_{\mathrm{PCB}}, T_{\mathrm{RCT(2:2)}}, T_{\mathrm{TCR(2:3)}}, T_{\mathrm{PSB}}, T_{\mathrm{CSB}}).$$

Považujeme-li součástky za nezávislé, funkci přežití minima lze přepsat jako součin funkcí přežití jednotlivých komponent.

Dosazením zjistíme, že

$$ET_{\mathrm{CELK}} \approx 43300 \text{ hodin} \approx 119 \text{ let}.$$

Odhad životnosti

Můžeme odhadnout:

- Střední dobu do poruchy pro jednotlivá zařízení – dosazením odhadnutých parametrů.
- Střední dobu do poruchy celého systému.

Díky redundanci usměrňovačů (stačí jeden ze dvou) a vyhodnocovacích jednotek TCR (dva ze tří) můžeme dobu do poruchy vyjádřit jako

$$T_{\mathrm{CELK}} = \min(T_{\mathrm{PCB}}, T_{\mathrm{RCT(2:2)}}, T_{\mathrm{TCR(2:3)}}, T_{\mathrm{PSB}}, T_{\mathrm{CSB}}).$$

Považujeme-li součástky za nezávislé, funkci přežití minima lze přepsat jako součin funkcí přežití jednotlivých komponent.

Dosazením zjistíme, že

$$ET_{\mathrm{CELK}} \approx 43300 \text{ hodin} \approx 119 \text{ let}.$$

To je šestkrát více, než požaduje norma, tedy je vše v pořádku.

Srovnání

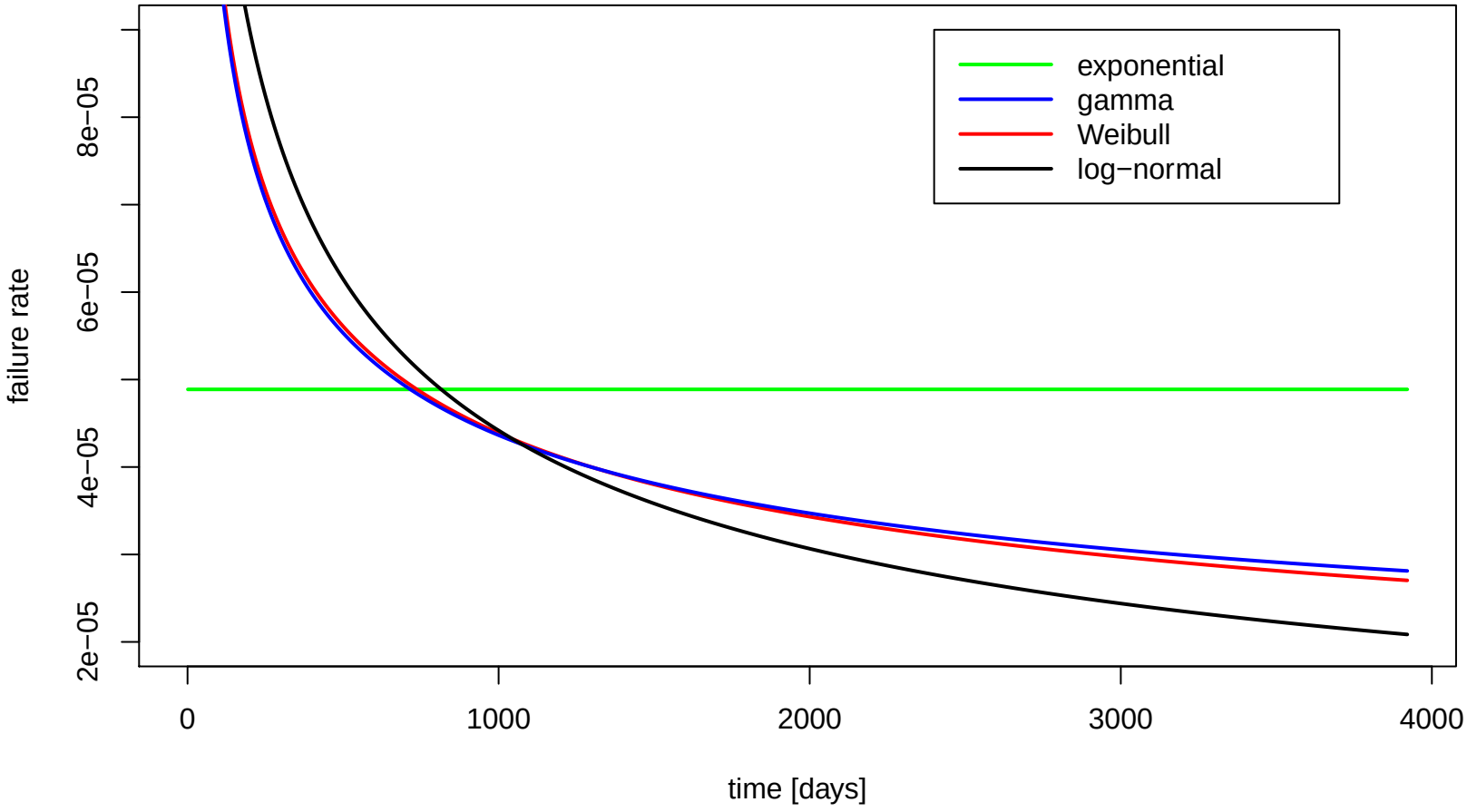
Siemens odhaduje střední dobu do poruchy detektoru Eurobalise ještě vyšší:

Trainguard Eurobalise S21 and S22

Technical data	
Speed range	0 to 500 kph (310 mph)
Power transmission frequency	27.095 MHz
Data transmission frequency	4.234 MHz
Data transmission rate	565 kbit/s
Type of modulation	Frequency shift keying (FSK)
Telegram length	341 bits or 1,023 bits (selectable)
Usable data length	210 bits or 830 bits (selectable)
Programming	contactless via an air gap by means of a test and programming unit
Operating distance LEU S21–Eurobalise	up to 5,000 m
Reliability (MTBF as per SN 29500)	S21 fixed-data balise > 800 years S21 transparent-data balise > 160 years

Zdroj: mobility.siemens.com

Odhadnuté intenzity poruch jednotky TCR



- Předpovídaná střední doba dožití leží mimo pozorovaný interval.
- Zkusíme neparametrický model pro bližší pohled na data.

Neparametrický odhad

- Pokud bychom znali rozprostření poruch a uvažovali bychom jen cenzorování zprava, funkci přežití bychom mohli odhadnout pomocí Kaplan-Meierova odhadu:

$$\hat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{d_j}{r_j}\right),$$

r_j – počet zařízení v riziku v čase T_j
d_j – počet zařízení, co se porouchaly v čase T_j.

Neparametrický odhad

- Pokud bychom znali rozprostření poruch a uvažovali bychom jen cenzorování zprava, funkci přežití bychom mohli odhadnout pomocí Kaplan-Meierova odhadu:

$$\hat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{d_j}{r_j}\right),$$

  r_j – počet zařízení v riziku v čase T_j
  d_j – počet zařízení, co se porouchaly v čase T_j.

- Zobecnění pro cenzorování zleva a intervalové cenzorování dle Turnbull (1976) a Huang and Wellner (1997) – iterativní postup.

Neparametrický odhad

- Pokud bychom znali rozprostření poruch a uvažovali bychom jen cenzorování zprava, funkci přežití bychom mohli odhadnout pomocí Kaplan-Meierova odhadu:

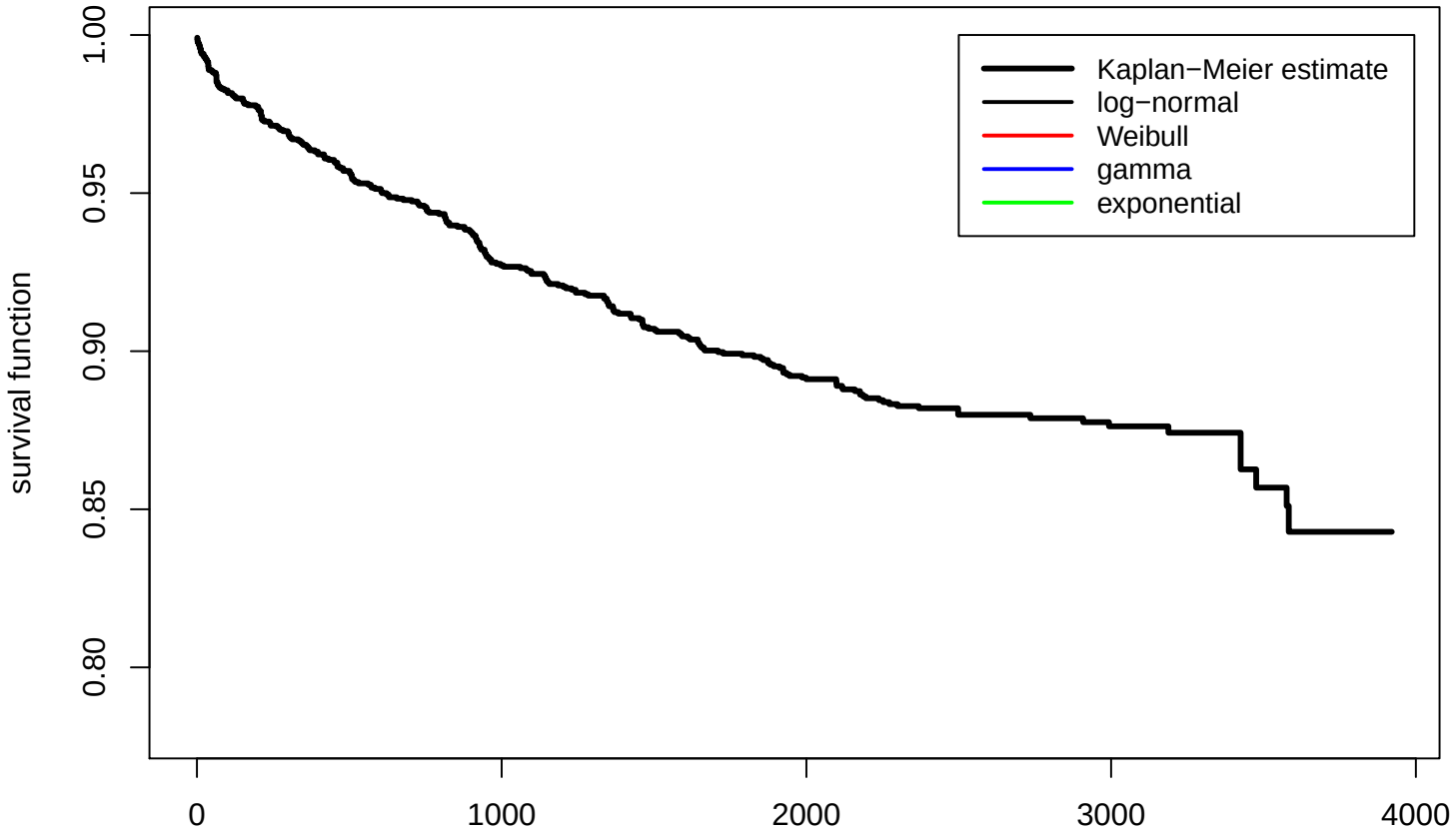
$$\hat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{d_j}{r_j}\right),$$

  r_j – počet zařízení v riziku v čase T_j
  d_j – počet zařízení, co se porouchaly v čase T_j.

- Zobecnění pro cenzorování zleva a intervalové cenzorování dle Turnbull (1976) a Huang and Wellner (1997) – iterativní postup.

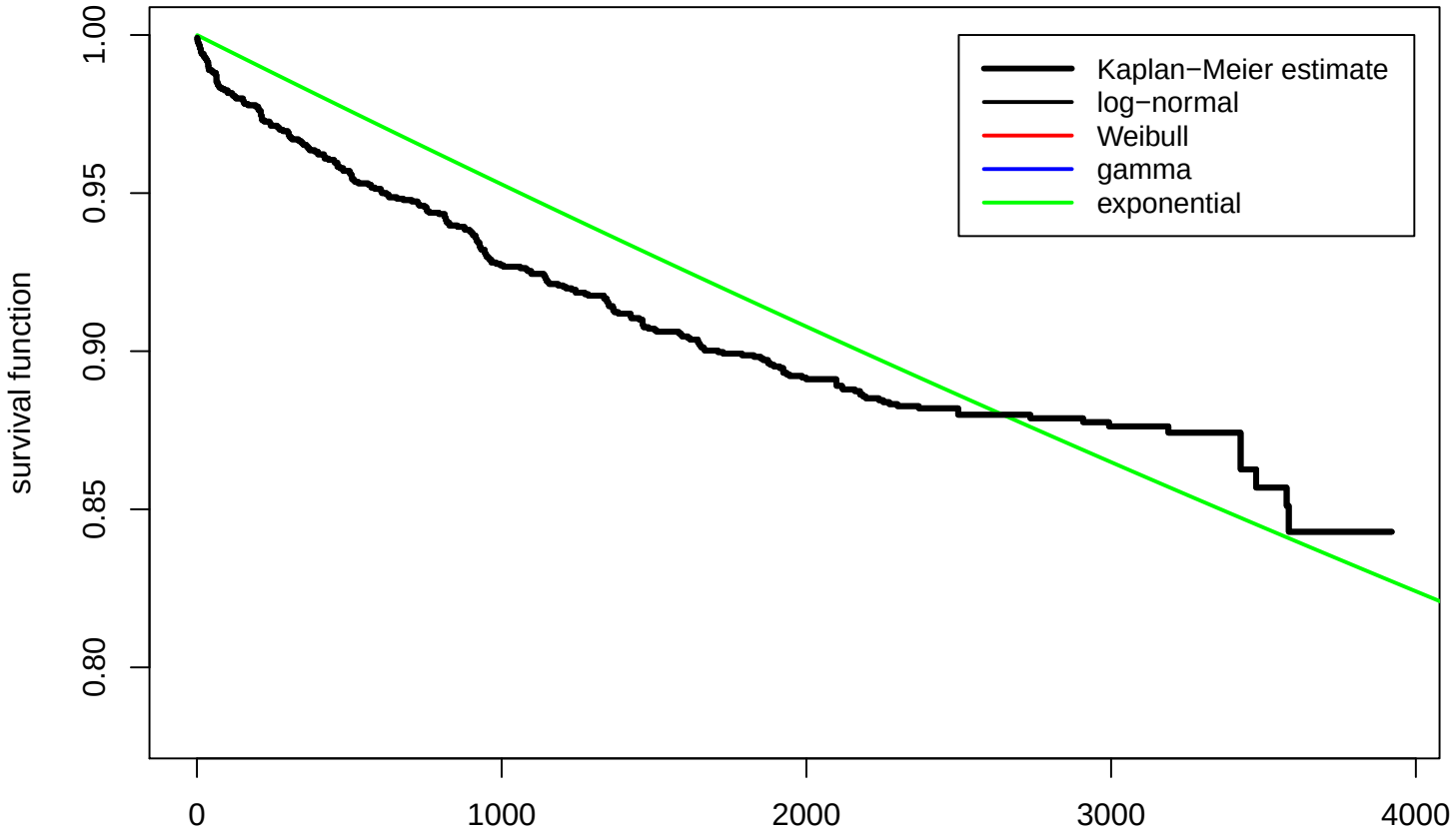
- Pokud předpokládáme neinformativní náhodné rozprostření poruch, můžeme jej vygenerovat, sestavit Kaplan-Meierův odhad a výsledek zprůměrovat. Získáme tak bodově konzistentní odhad.

Neparametrický odhad funkce přežití pro detektor TCR



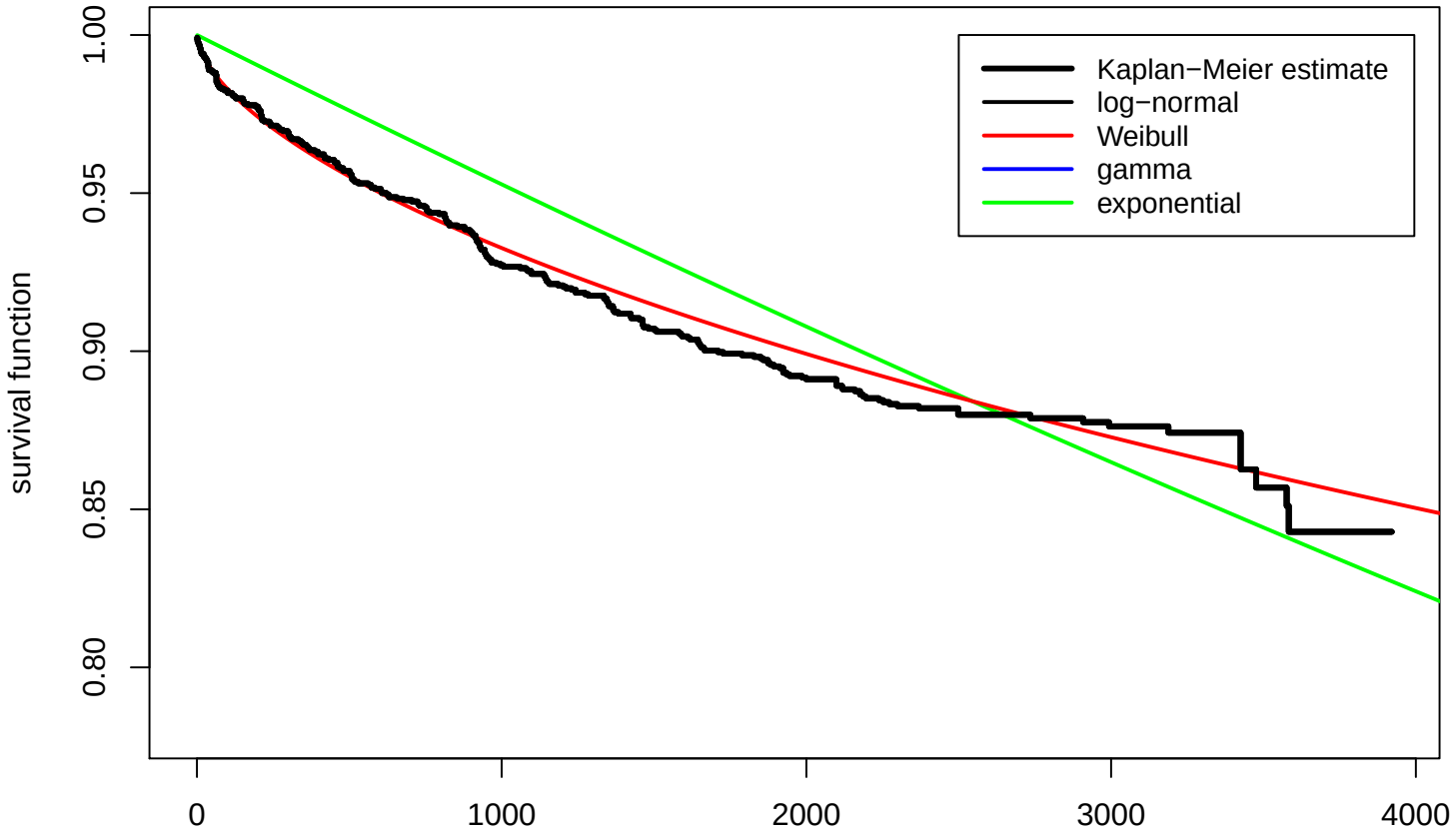
- Weibullovo, gamma a log-normální rozdělení popisují dobře chování dat na začátku intervalu.
- Ke konci intervalu nesrovnalosti. Možný náznak vanové křivky.

Neparametrický odhad funkce přežití pro detektor TCR



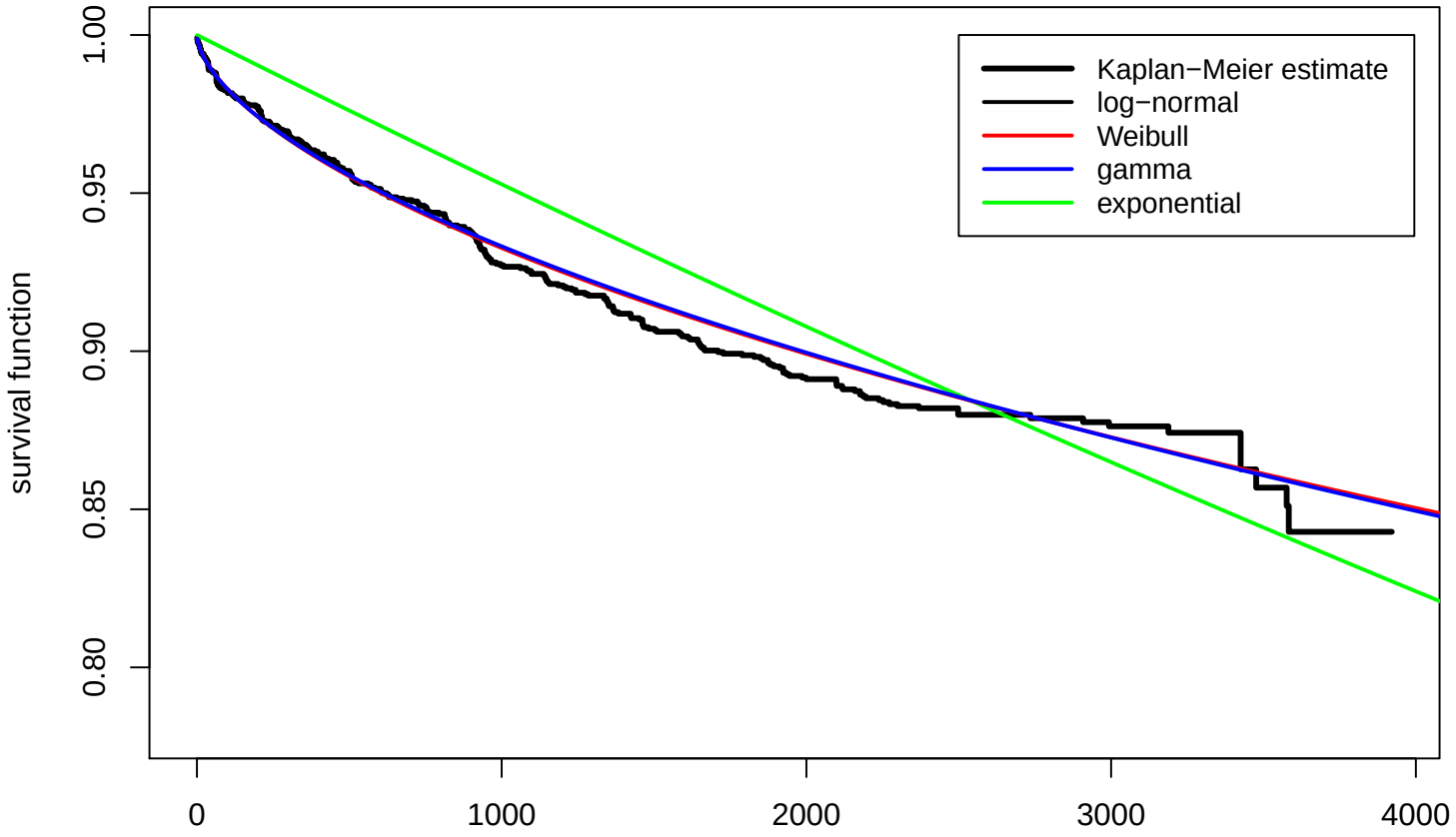
- Weibullovo, gamma a log-normální rozdělení popisují dobře chování dat na začátku intervalu.
- Ke konci intervalu nesrovnalosti. Možný náznak vanové křivky.

Neparametrický odhad funkce přežití pro detektor TCR



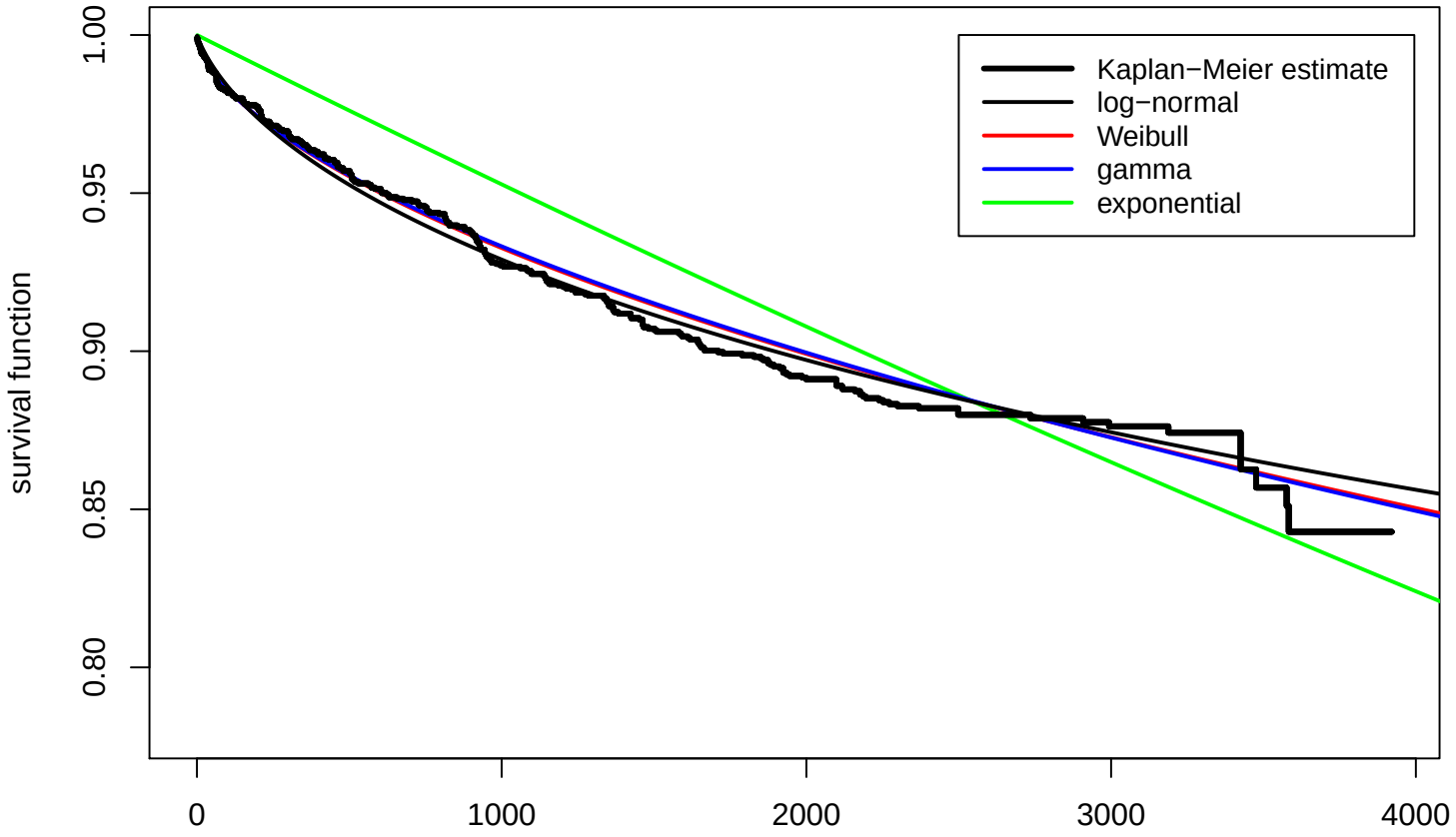
- Weibullovo, gamma a log-normální rozdělení popisují dobře chování dat na začátku intervalu.
- Ke konci intervalu nesrovnalosti. Možný náznak vanové křivky.

Neparametrický odhad funkce přežití pro detektor TCR



- Weibullovo, gamma a log-normální rozdělení popisují dobře chování dat na začátku intervalu.
- Ke konci intervalu nesrovnalosti. Možný náznak vanové křivky.

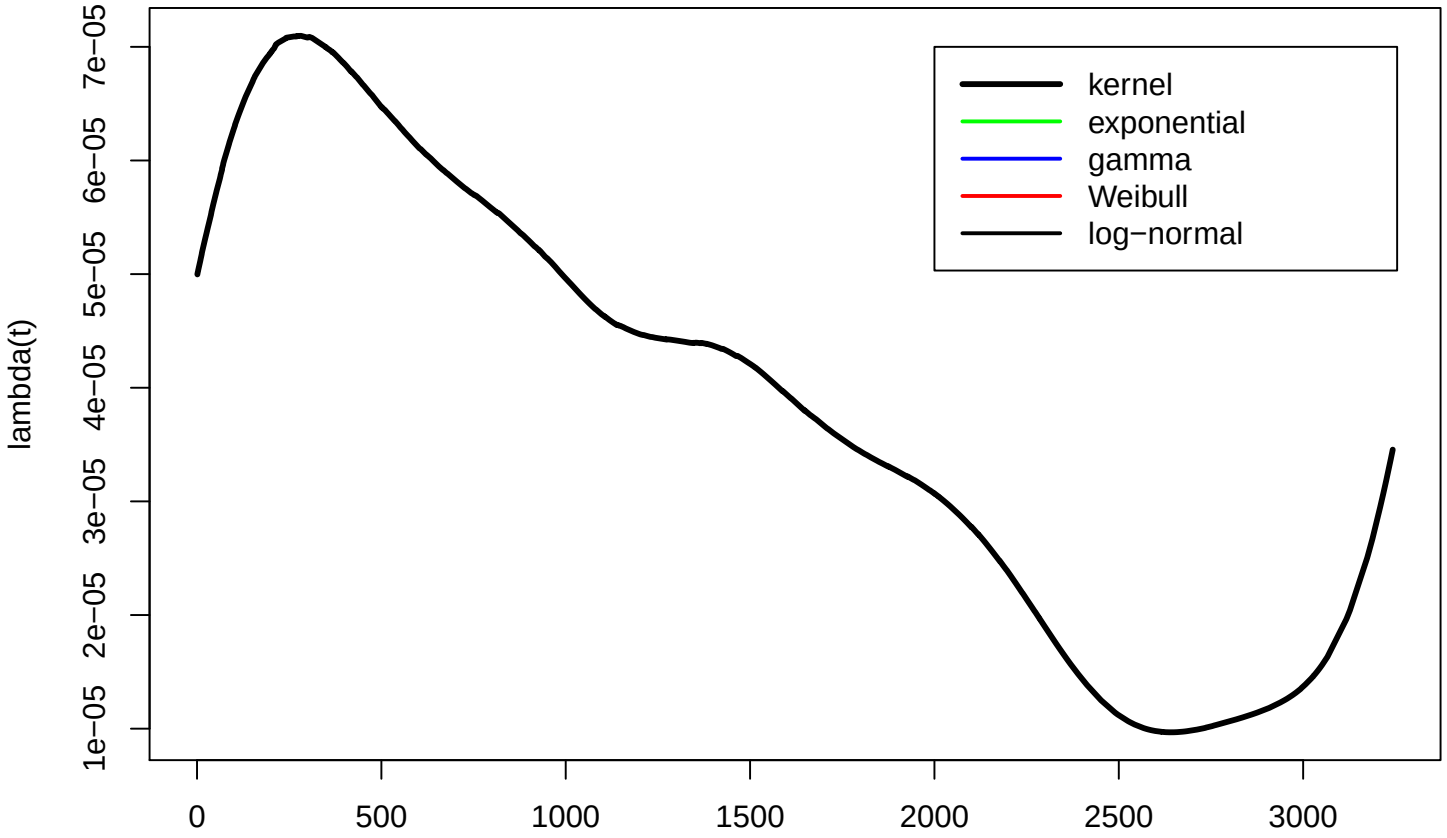
Neparametrický odhad funkce přežití pro detektor TCR



- Weibullovo, gamma a log-normální rozdělení popisují dobře chování dat na začátku intervalu.
- Ke konci intervalu nesrovnalosti. Možný náznak vanové křivky.

Odhadnuté intenzity poruch jednotky TCR

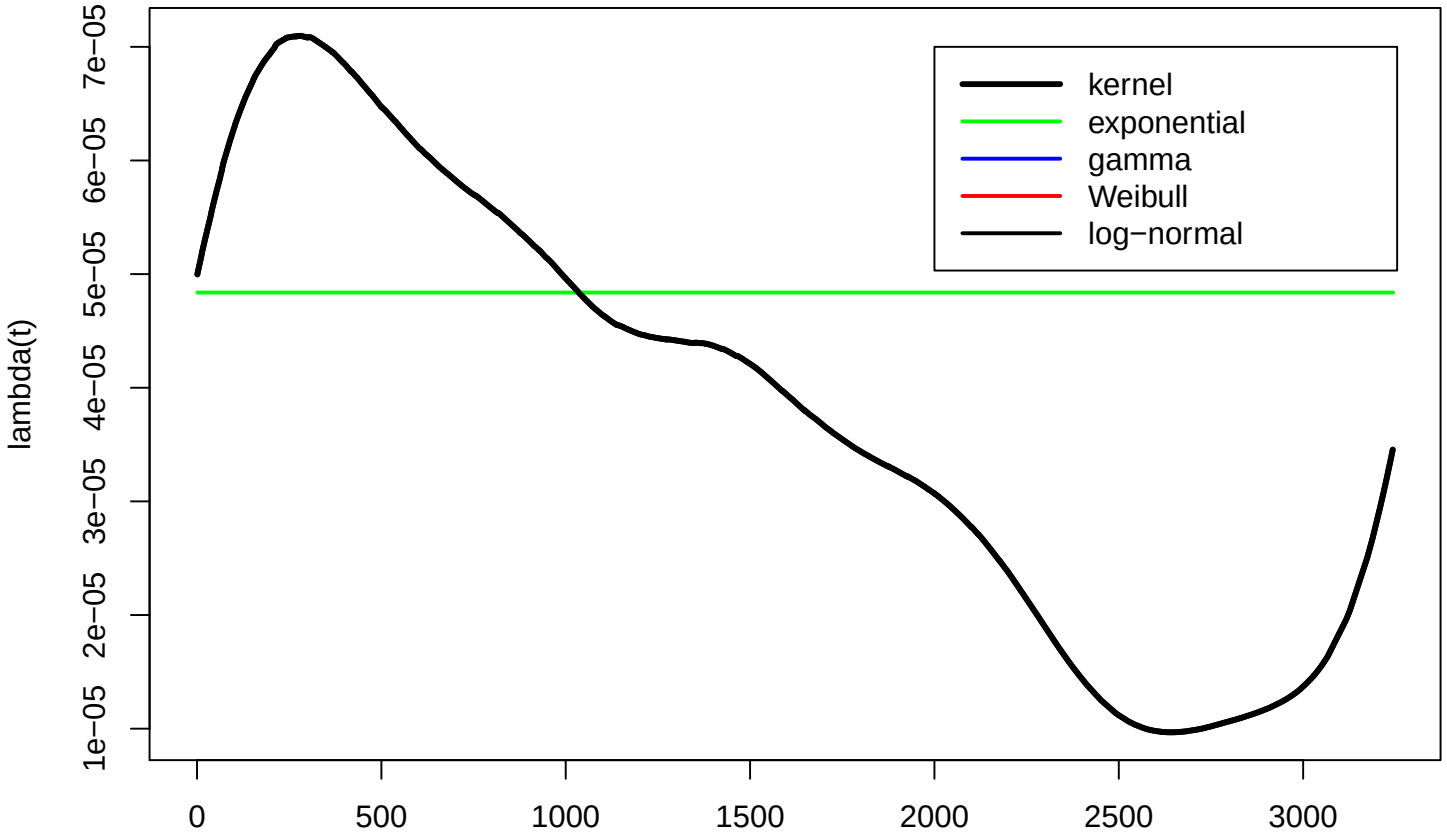
Odhadneme intenzitu poruch pomocí jádrových odhadů (Diehl and Stute, 1988) – zjednodušeně, pro jedno konkrétní rozřazení poruch:



Ke konci pozorovaného intervalu je skutečně nesrovnalost.

Odhadnuté intenzity poruch jednotky TCR

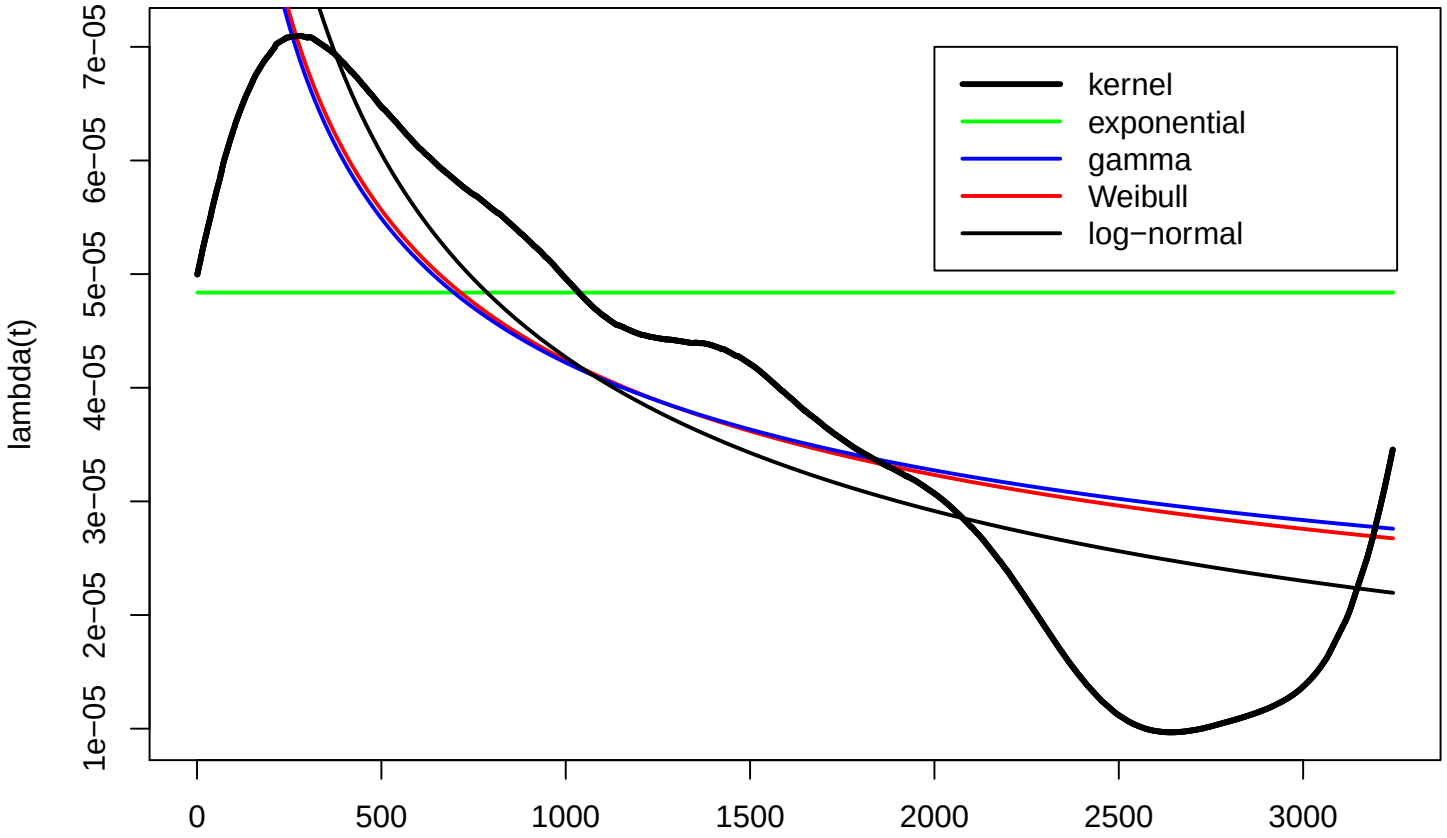
Odhadneme intenzitu poruch pomocí jádrových odhadů (Diehl and Stute, 1988) – zjednodušeně, pro jedno konkrétní rozřazení poruch:



Ke konci pozorovaného intervalu je skutečně nesrovnalost.

Odhadnuté intenzity poruch jednotky TCR

Odhadneme intenzitu poruch pomocí jádrových odhadů (Diehl and Stute, 1988) – zjednodušeně, pro jedno konkrétní rozřazení poruch:



Ke konci pozorovaného intervalu je skutečně nesrovnalost.

Závěr

Shrnutí:

- Zkoumali jsme metody, jak na základě dat z 11 let provozu modelovat rozdělení doby do poruchy elektronických kolejových obvodů.
- Použili jsme parametrické i neparametrické metody pro cenzorovaná data.
- Museli jsme řešit nekompletnosti v datech, zejména chybějící přiřazení poruch k zařízením.
- Odhad střední doby do poruchy není dobře použitelný.

Závěr

Shrnutí:

- Zkoumali jsme metody, jak na základě dat z 11 let provozu modelovat rozdělení doby do poruchy elektronických kolejových obvodů.
- Použili jsme parametrické i neparametrické metody pro cenzorovaná data.
- Museli jsme řešit nekompletnosti v datech, zejména chybějící přiřazení poruch k zařízením.
- Odhad střední doby do poruchy není dobře použitelný.

Outlook:

- Prozkoumat další vlastnosti odhadů – intervaly spolehlivosti pro Kaplan-Meierův odhad apod.
- Vyzkoušet modely neúplné opravy a konkurujících rizik, případně regresní modely (Cox, 1972).
- Prošetřit chování dat ke konci intervalu.

Reference

[1] Cox D.R.: *Regression models and life tables*. J. Roy. Statist. Soc. Ser. B 34, 187-220, 1972.

[2] Diehl S., Stute W.: *Kernel density and hazard function estimation in the presence of censoring*. Journal of Multivariate Analysis 25(2), 299-310, 1988.

[3] Huang J., Wellner J.A.: *Interval Censored Survival Data: A Review of Recent Progress*. In: Lin D.Y., Fleming T.R. (eds) Proceedings of the First Seattle Symposium in Biostatistics. Lecture Notes in Statistics, vol 123. Springer, New York, NY, 1997.

[4] Turnbull, B.W.: *The empirical distribution function with arbitrarily grouped, censored and truncated data*. J. Roy. Statist. Soc. B 38, 290–295, 1976.

[5] Siemens AG: *Trainguard Eurobalise S21 and S22* at `https://www.mobility.siemens.com/mobility/global/SiteCollectionDocuments/en/rail-solutions/rail-automation/train-control-systems/trainguard-eurobalise-s21-s22-en.pdf` [2018-01-22].