Analýza historie událostí (event history analýza) – možnosti a základní principy při studiu životních drah*

Anna Šťastná**
Výzkumný ústav práce a sociálních věcí, v. v. i.

Event history analysis - possibilities and the basic principles for study life-courses

Abstract: *Event history modelling techniques have become increasingly widespread in the social sciences over the last few decades and the range of applications includes demographic and sociological analyses, labour market studies, mobility and migration studies, as well as analyses within political science. In principle, event history analysis represents an extension of the statistical techniques connected with the life table method and can be defined as an analysis of the duration of the non-occurrence of a given event during a risk period. This article devotes attention to the concept of event history analysis in terms of data considerations, basic principles and methods of analysis. In order to discuss the basic methods and their potential to interpret results, the author applied the event-history approach to an analysis of the process of leaving the parental home using data from the Czech Generations and Gender Survey [2005]. The final part of this study discusses some key issues involved in using the event history approach when analysing socio-demographic topics within the Czech context.*

Key words: *event history analysis, life-course, life-history data, transition to adulthood.*

Data a výzkum - SDA Info 2011, Vol. 5, No. 1: 59-83.

* Tento text vznikl v rámci grantového projektu MŠMT „Aktivní stárnutí, rodina a mezigenerační solidarita“ (č. 2D06004).
** Veškerou korespondenci posílejte na adresu: PhDr. Anna Šťastná, Výzkumný ústav práce a sociálních věcí, v. v. i., Palackého nám. 4, 12801 Praha 2; www.vupsv.cz.

V posledních desetiletích se ve společenských vědách rozvíjejí specifické statistické metody, které se zaměřují na výzkum životní dráhy jedince. Tyto postupy umožňují sledovat časování základních životních událostí a studovat faktory, které výskyt daných událostí ovlivňují. Činí tak prostřednictvím výpočtu rizika, že u jedince nastane sledovaná událost. Tyto metody bývají v sociálních vědách označovány nejčastěji jako *event history analysis* a s jejich aplikací se setkáváme například v demografických a sociologických analýzách, studiích trhu práce stejně tak jako ve zkoumání mobility i migrace. Tento soubor metod se přitom rozvíjel nezávisle v rámci různých vědních oborů, proto se setkáme také se škálou jejich označení: v medicíně a epidemiologii se nejčastěji používá analýza přežívání (*survival analysis*), v ekonomii analýza času trvání (*duration analysis*), v technických vědách se pak užívá označení analýza spolehlivosti, resp. poruchovosti (analýza časů selhání – *failure time analysis*, *reliability analysis*).

Tento text se zaměří na koncepci analýzy historie událostí, její základní principy, metody a nezbytná data včetně ukázky základních typů výstupů a jejich interpretačních možností. Dále jsou uvedeny oblasti, ve kterých je možné analýzy historie událostí aplikovat, základní učební a přehledové texty a také nezbytný software.

1. Životní dráha a její studium jako východisko analýzy historie událostí

Na životní dráhu pohlížíme jako na proces vývoje od dětství přes adolescenci a dospělost do zralého věku a stáří [Alan 1989]. Cílem analýzy životních drah je popsat, vysvětlit a předpovědět hlavní události, které lidé prožívají. Studium se tedy zaměřuje na odhalování vzorců v časování a následnosti životních událostí, na určení, zda a jak jsou jednotlivé životní události spojeny, a na předpověď nebo rekonstrukci životních drah z parciálních pozorování [Willekens 1999]. Koncepce životní dráhy umožňuje propojit řadu procesů, které bývají při analýzách sledovány relativně autonomně.

Životní historie může být popisována v termínech životních událostí nebo stádií mezi událostmi. Lidé mají různé životní dráhy a jejich odlišnosti spočívají především v typech vyskytujících se událostí, v počtu událostí, v jejich časování a posloupnosti. Rozdíly v biografiích vysvětluje mnoho faktorů, některé mohou být dány chováním, jiné jsou biologické povahy. Některé mohu být spojovány s hodnotami a aspiracemi jedince, jiné mohou záviset na sociálních a kulturních normách, právních omezeních a přístupu k ekonomickým i jiným zdrojům.

Při studiu životních drah založeném na analýze kvantitativních longitudinálních a panelových dat se v posledních desetiletích v sociálních vědách využívá řada metodických postupů [více např. Pakosta, Fučík 2009; Singer, Willett 2003]. Analýzy sledující životní trajektorie jedinců jsou založené na sekvenční analýze a sledování posloupnosti stavů, ze kterých se v určitém časovém období sestává životní trajektorie jedince [česky viz Chaloupková 2009, 2010]. V analýzách, kde je vedle časového rozměru nutné respektovat také hierarchickou povahu dat, je možné využít víceúrovňové modelování [česky viz Pakosta, Fučík 2009].

Hojně využívaným přístupem je analýza historie událostí (event history analýza), které je věnován tento text. Jedná se o statistické metody studující výskyt jednotlivých událostí a jejich načasování. Lze je chápat jako analýzu doby, kdy je jedinec vystaven možnosti (riziku) výskytu události a po kterou k této události nedojde [Yamaguchi 1991]. V rámci konceptu studia životní dráhy se tyto modely zaměřují na životní události, tj. „události, které znamenají změnu z jedné životní fáze (etapy) do druhé“ [Willekens 1999: 44]. Měření času, po který je jedinec vystaven riziku výskytu události, je pro tyto modely zásadní. Období vystavení riziku je možné definovat jako epizodu v životě, která je zahájena jednou životní událostí a ukončena událostí jinou. Hlavní sledovaná proměnná je tedy čas výskytu nějaké události, resp. doba do určité události [Hendl 2004: 44].

Principy analýzy historie událostí se vyvinuly ze souboru metod spojených s tabulkami života tradičně využívanými v demografii [Hoem 1993 citován in Manting 1994, více např. Rychtaříková 2008]. Princip tabulek života byl posléze rozpracován a pro metody event history analýzy byly zavedeny také modely založené na principech regrese.

Je nutné předeslat, že metody a přístupy k analýze historie událostí lze rozdělit na dvě velké skupiny dle toho, zda je čas, který je klíčovou proměnnou ve všech těchto modelech, měřen jako spojitá, nebo jako diskrétní proměnná. Celý následující text je věnován případu, že je s časem pracováno jako se spojitou proměnnou. Stručně se však na tomto místě zmíníme o modelech s diskrétním časem, neboť v takovémto případě je třeba přistoupit k jiné analytické strategii. Vždy je vhodné měřit čas pokud možno v co nejpřesnějších časových jednotkách, v řadě případů to však není možné (především v retrospektivně sbíraných datech), případně diskrétní povaha proměnné vyplývá již ze samotné povahy zjišťovaných dat – některé události se mohou vyskytnout pouze v diskrétním čase (např. středoškolští studenti mohou úspěšně dokončit studium pouze v několika málo – diskrétních – okamžicích v průběhu roku). Rozlišování mezi časem jako spojitou a diskrétní proměnnou není pouhým metodologickým detailem, neboť téměř všechny vlastnosti analýzy výskytu událostí (definice parametrů, konstrukce modelů, odhady i jejich testování) záleží právě na tomto rozlišení [Singer, Willett 2003]. Modely, které pracují s časem v jeho nespojité podobě, jsou *discrete-time modely* a v tomto textu jim nebude věnována pozornost. Více o těchto modelech je možné nastudovat v příslušných publikacích [např. Allison 1982; Singer, Willett 2003; Vermunt 1997a; Yamaguchi 1991].

2. Základní pojmy a principy analýzy historie událostí

Základními jednotkami životní dráhy jsou životní události a každá takováto událost má čtyři aspekty: typ události, dobu výskytu, pravděpodobnost výskytu (riziko, že událost nastane) a důvod výskytu (vliv ostatních událostí a procesů). Jedna z hlavních charakteristik analýzy historie událostí je její implicitní důraz na čas – na plynutí individuálního času, na individuální změny vyskytující se spolu s časem i na historické časové souvislosti, ve kterých se individuální životní dráha odehrává.

Při úvahách o tom, zda je pro studovanou problematiku vhodné použít právě metody analýzy historie událostí, doporučuje Singer a Willett [2003: 306] použít jednoduchý test („whether and when test"), který spočívá v položení si výzkumných otázek ohledně studovaného tématu, a pakliže obsahují otázku „zda?" a „kdy?", pak se zřejmě jedná o oblast, kde by se hodilo využití následujících technik. Tyto metody se používají tehdy, pokud jsou při studiu procesu kladeny otázky typu: Kdy k události dochází? Jak se určité kategorie či skupiny podílejí na rozdílech v časování události? Jaké jsou rozdíly ve výskytu události v různých obdobích, v různých sociálních kategoriích, mezi zeměmi? Smyslem je tedy vysvětlit, kteří jedinci a proč jsou vystaveni vyššímu nebo naopak nižšímu riziku podstoupit studovanou událost než jiní. Klíčovými koncepty tohoto analytického přístupu jsou tedy **čas** a **riziko**.

V diskusích ohledně chápání času a jeho studia v rámci zkoumání životních drah navrhují někteří autoři rozlišovat mezi časem chronologickým a vývojovým[1]. **Chronologický čas** lokalizuje události na zvolené časové stupnici, použito přitom může být několik časových stupnic (např. kalendářní čas, věk, čas uplynulý od určité události). Běžně se rozlišuje mezi věkem (individuální čas), trváním procesu (procesní čas) a historickým časem (kalendářní čas). Časové škály mohou být pojímány jako hodiny, které začínají měřit čas v různých počátcích. K vyjádření času pak lze použít různé stupnice měření.

Měření času zároveň přináší řadu metodologických problémů. Zmiňme se o dvou: (1) Čas, kdy se událost v životě jedince vyskytne, se v řadě případů odlišuje od času, kdy je její výskyt iniciován. V ideálním případě bychom měli uvažovat čas začátku procesu, ten však mnohdy není možné přesně určit. Například studium plodnosti často uvažuje vliv vybraných faktorů na časování narození dítěte, ačkoliv by v tomto případě bylo lépe uvažovat vliv na časování koncepce (začátek těhotenství[2]). Také např. rozpad manželství je dlouhodobějším procesem a datem rozvodu je pak zpravidla pouze završen. Právě okamžik rozvodu je však možné jasně operacionalizovat a dotazovat se na jeho dataci ve výzkumném šetření. (2) Druhý problém nastává v případě, že se měřicí stupnice používaná badatelem odlišuje od škály používané respondentem. Rozdíl spočívá ve vztahu mezi objektivní měřitelností a subjektivním vnímáním času a v napětí mezi časem „externím", tj. daným, a „interním", tj. prožívaným časem [Willekens 1999:

1 V*ývojový čas* situuje události v procesu fyziologického vývoje. Např. Sinclair [cit. dle Willekens 1999: 33] se zabýval několika znaky vyspělosti (rozlišoval věk kostí – radiologický věk, věk zubů, sexuální věk, nervový, mentální i psychologický věk) a došel k závěru, že chronologický věk, obecně užívaný při studiu životní dráhy, by měl být spíše nahrazen časem vývojovým. Vývojový čas se více vztahuje k míře a rychlosti změn v dané fázi života či k počtu životních událostí podstoupených do té doby. Operacionalizace takto chápaného času je však pro analýzy založené na datech výběrových šetření téměř vyloučená.

2 V analytických modelech je možné tuto situaci řešit např. tak, že studovanou událostí zůstává narození dítěte, tato událost je však namísto data narození měřena v době početí, která je schematicky odhadnuta jako doba 9 měsíců před datem narození, které je respondentem uvedeno v dotazníku.

35]. Astronomický čas je uniformní, čistě kvantitativní a kontinuální, naproti tomu *sociální čas* má kvalitativní charakter, neplyne rovnoměrně a není libovolně dělitelný [Šubrt 1993]. Nepřesné výpovědi respondentů ohledně doby či věku, kdy zažili danou událost, nebo o délce určité životní etapy mohou pramenit z různého vnímání času a také ze sklonu lidí zaokrouhlovat či řadit události, u nichž si nejsou přesně jisti dobou výskytu, spíše k zokrouhlenému datu či věku (často končícímu na číslice 0 nebo 5). Ve výsledku se to projeví abnormální koncentrací výskytu událostí k určitému datu nebo délce trvání (můžeme se s tím setkat např. při studiu historických populací). Mylné udání datace události může být z části ovlivněno také vnímáním životních událostí nikoli v čase chronologickém, ale ve vzájemných relacích jedněch událostí s druhými.

Jedinec je vystaven **riziku** výskytu události pouze tehdy, pakliže může studovanou událost podstoupit (například pouze tzv. sňatkuschopné obyvatelstvo může uzavírat manželství, pouze vdané a ženatí jsou vystaveni riziku rozvodu, pouze plodné ženy mohou počít). Riziko výskytu události tedy znamená, že zde existuje možnost, že u sledovaného jedince nastane studovaná událost. Lidé se obecně liší podle úrovně podstoupeného rizika a doby, po kterou jsou tomuto riziku vystaveni. Indikátory rizika jsou v analýze historie událostí klíčové, neboť jsou v těchto modelech závislými proměnnými. V modelech je sledováno riziko výskytu studované události v daném období a mezi jednotlivými sledovanými skupinami.

Vedle těchto dvou klíčových pojmů metodologický koncept analýzy historie událostí rozlišuje a jasně definuje několik dalších pojmů. Při analýze daného jevu je nutné definovat tzv. **výchozí událost**, kterou podstoupili všichni dále studovaní jedinci a která je definována jednoznačně a jasně (např. při studiu rozvodu prvního manželství je výchozí událostí první sňatek; následná analýza se odehrává na souboru těch, kteří někdy uzavřeli manželství). Dále je nutné definovat tzv. **studovanou událost**, která je vždy dána formulací problému (zde tedy rozvod prvního manželství). Studovanou událost chápeme jako kvalitativní změnu, kterou lze situovat v čase. Změna by měla znamenat relativně ostrý předěl mezi tím, co předcházelo a co následovalo [Allison 1984: 9]. Životní události mohou být neopakovatelné povahy (mohou nastat pouze jednou za život), nebo se mohou v průběhu životního cyklu opakovat, resp. nastat několikrát (např. sňatek, narození dítěte, nezaměstnanost). Pro potřeby analýzy historie událostí je nutné studovanou událost definovat jednoznačně, a to i v případě událostí svým charakterem opakovatelných. V takovýchto případech (např. narození dítěte) obvykle studovanou událost jasně specifikujeme pomocí odlišení pořadí události (narození 1. dítěte, 2. dítěte apod.).

Všichni jedinci, kteří podstoupili výchozí událost a u kterých tedy může nastat studovaná událost, tvoří tzv. **risk set**, populaci vystavenou riziku výskytu události. Tento soubor je v čase proměnlivý spolu s tím, jak jeho jednotliví členové postupně studovanou událost podstupují, případně ze sledovaného souboru vypadávají (např. při studiu sňatečnosti na počátku sledujeme soubor 1 000 svobodných žen, po měsíci sledování 4 ženy uzavřely sňatek, 2. měsíc je tedy risk set o tento počet snížen, neboť sňatek může uzavřít již pouze 996 žen).

Na základě toho, zda jedinec podstoupil či nepodstoupil studovanou událost, je vytvořena nová pomocná proměnná, která je indikátorem **cenzorování**. Možnost zahrnout do analýz i pozorování, která jsou nějakým způsobem cenzorovaná, patří k jedné z velkých výhod těchto analytických postupů [Yamaguchi 1991; Manting 1994; Lelièvre, Bringé 1998]. Nejčastějším případem je tzv. *cenzorování zprava*,[3] kdy jedinec po celou dobu sledování studovanou událost nepodstoupil. Pozorování jsou cenzorována v případech, kdy po dobu sledování nedošlo ke studované události (např. ne všem se narodí do okamžiku šetření dítě, bezdětná respondentka je tedy cenzorována okamžikem šetření), ale také v případech, kdy jedinec přestal být vystaven riziku výskytu události (vypadl z risk setu) z nějakého jiného důvodu, než je datum šetření – např. z důvodu jiné, rušivé události, jako je emigrace sledovaného jedince. Pozorování může být cenzorováno také stanovenou věkovou hranicí, v případě porodnosti např. hranicí plodného věku ženy, při studiu rozvodovosti je pozorování cenzorováno také v okamžiku ovdovění jedince apod. Tato pomocná proměnná je binární, v nejčastěji používaných statistických programech (např. SAS, SPSS, STATA) nabývá hodnoty 1 v případě, že u daného respondenta studovaná událost nastala, a hodnoty 0 v případě, kdy bylo pozorování cenzorované.

Klíčovou proměnnou v tomto typu analytických modelů je časová proměnná označující délku **sledované časové epizody** (*process time/duration*). Jinými slovy jedná se o období, po které je jedinec vystaven riziku prožít studovanou událost, tzv. risk period. Při měření časové epizody, po kterou je jedinec vystaven možnosti výskytu události, je třeba: (a) chápat expozici jako epizodu v životní dráze, která je iniciována výskytem jedné události a ukončena výskytem události jiné (např. doba, po kterou je jedinec vystaven riziku rozpadu manželství, začíná okamžikem sňatku a je ukončena rozvodem nebo úmrtím jednoho z partnerů); (b) rozlišovat mezi trváním expozice a délkou pozorování osob během výzkumu. Ve výzkumech často není možné pozorovat osoby během celého období, po které jsou vystaveni riziku výskytu studované události, dostupná data totiž pokrývají pouze část takovéhoto období[4]. Tato časová proměnná, která je v analýze historie událostí závislou proměnou, je konstruována následovně:

- podstoupil-li jedinec studovanou událost, je hodnota časové proměnné rovna době mezi výchozí událostí a studovanou událostí (např. při studiu rozvodovosti byl jedinec vystaven riziku rozvodu od data sňatku – výchozí událost, až do okamžiku samotného rozvodu);
- nenastala-li studovaná událost a pozorování bylo cenzorované, hodnota časové proměnné je rovna době mezi výchozí událostí a okamžikem cenzorování (v uvedeném případě byl jedinec vystaven riziku rozvodu od data sňatku až do okamžiku konce šetření, případně do okamžiku, kdy ovdověl).

3 Existuje například i tzv. *cenzorování zleva*, které se vyskytuje v případě, že chybí informace o vstupu do určité epizody – např. při studiu procesu rozvodovosti chybí u sezdaného jedince informace o době uzavření sňatku [Manting 1994].

4 Velké množství výzkumů je navíc prováděno retrospektivně a délka vystavení respondenta riziku zkoumané události je tak zjišťována a vypočítávána zpětně.

Cílem analýzy historie událostí není pouze popsat výskyt události a její časování, ale postihnout také faktory, které výskyt a časování ovlivňují. Cílem je tedy zahrnout do modelu **sadu vysvětlujících proměnných**. Vedle proměnných, které v čase nemění svoji hodnotu (*time-constant/fixed covariates*), jako například rok narození respondenta či jeho pohlaví, dovoluje řada modelů zahrnout také vysvětlující proměnné, které mohou během sledovaného období nabývat různých hodnot (*time-varying covariates*). Příkladem takovéto proměnné může být nejvyšší ukončené vzdělání, rodinný stav, počet dětí či velikost místa bydliště/region/země. U některých vysvětlujících proměnných je jejich charakter zřejmý a jednoznačný (datum narození respondenta je jednou pro vždy dané), u jiných však záleží na kontextu a výzkumné otázce. To, zda je vysvětlující proměnná v čase neměnná, nebo se naopak proměňuje, může totiž záviset i na studovaném procesu. Při studiu narození prvního dítěte, kdy sledujeme ženy například od jejich 15 let věku, se výše jejich vzdělání v řadě případů výrazně proměňuje spolu s rostoucím věkem.[5] Studujeme-li však například u skupiny osob nad 55 let proces přechodu do důchodu, je vhodnější zahrnout nejvyšší ukončené vzdělání jako vysvětlující proměnnou konstantní v čase. V tomto případě není příliš relevantní zahrnovat vzdělání formou měnící se proměnné, neboť své formální vzdělání v daném věku zvyšuje pouze minimum osob, které, i když se je podaří zachytit v datovém souboru, nepokryjí dostatečně dané kategorie.

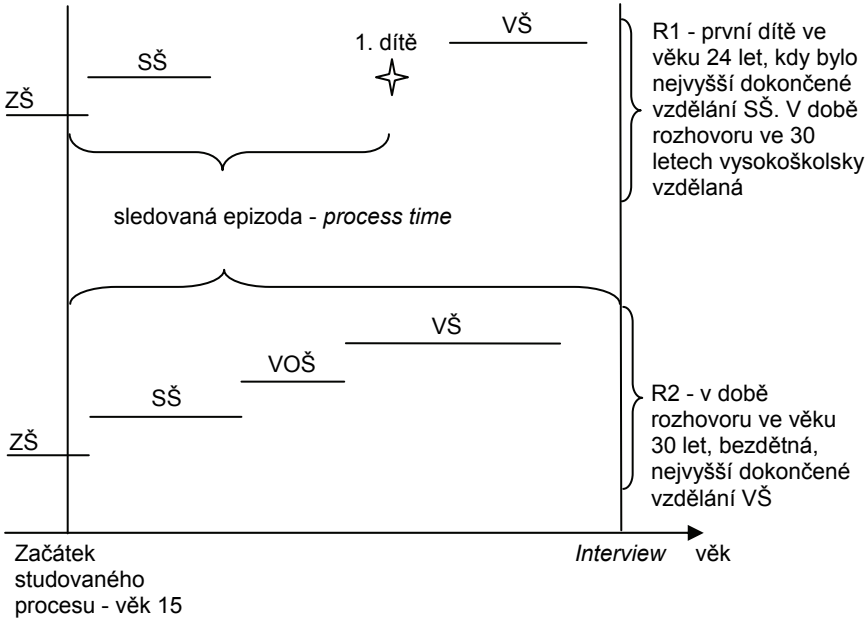
Shrneme-li základní principy a pojmy graficky, pak Obrázek 1 (viz následující stranu) znázorňuje příklad dvou respondentek, u kterých sledujeme proces vstupu do rodičovství – výchozí událostí jsou 15. narozeniny ženy, studovanou událostí narození 1. dítěte. Za obě respondentky (R1 a R2) jsou k dispozici informace do 30 let věku, sledovaná epizoda (*process time*) a tedy časová proměnná vstupující do analytického modelu se však různí – u R1 je ukončena studovanou událostí, u doposud bezdětné R2 až okamžikem rozhovoru (pozorování je cenzorované). Obrázek zároveň zachycuje jednu vysvětlující proměnnou – vzdělání – a její proměny v průběhu sledované epizody.

Výše popsané principy naznačují důvody, pro které nejde použít v těchto analýzách klasické popisné statistiky a regresní přístupy. Okolnost, že některá pozorování v souboru jsou cenzorovaná, představuje problém pro klasickou jednorozměrnou popisnou statistiku, jako je průměr, modus či medián. Logistické regresní modely mohou analyzovat, zda událost nastala či nikoli, nepracují však s časem jako závisle proměnnou (víme tedy, že jedinci se do 30. roku věku narodilo dítě, model však nebere v úvahu, zda k tomu došlo v 19 či 29 letech). V případě lineární regrese představuje vedle cenzorování problém především předpoklad normálního rozdělení. Rozdělení časů přežití v event history analýze je téměř vždy asymetrické, může být bimodální, a na tyto odchylky není lineární regrese dostatečně robustní[6] [Cleves, Gould, Gutierrez, Marchenko 2010]. Sa-

5 K tzv. anticipatornímu modelování a zahrnutí v čase se měnících proměnných při analýze plodnosti více viz Hoem, Kreyenfeld [2006].

6 O odhadu hovoříme jako o robustním, pokud není příliš citlivý k odchylkám od předpokládaného rozdělení.

Obrázek 1.



mostatným specifikem, které dovoluje analyzovat modely event history analýzy, jsou vysvětlující proměnné, které se mění v čase.

3. Life-history data – požadavky na formát dat a datové zdroje

Použití vysvětlujících proměnných, které v průběhu sledované epizody nabývají různých hodnot, si vyžaduje specifické zacházení s datovým souborem. Při něm přestává být jednotkou respondent a důraz je kladen na jednotlivé časové úseky, ve kterých nabývá vysvětlující proměnná různých hodnot. Pro jednoho respondenta tedy již nemusí být v datové matici určen pouze jeden řádek, ale epizoda, po kterou je tento jedinec vystaven riziku výskytu události, je rozdělena na více časových intervalů (a tedy řádků), jejichž počet závisí na tom, kolika hodnot nabývají vysvětlující proměnné. Mezi jednotlivými respondenty se tedy může lišit. Ilustrujme podobu datová matice na uvedeném příkladě analýzy narození 1. dítěte.

Tabulka 1 znázorňuje dvě respondentky a případ, kdy je vysvětlující proměnná fixní (rok narození). První sloupec značí identifikační číslo respondenta, druhý („TIME“) značí délku sledované epizody (*process time*), po kterou je respondentka vystavena riziku výskytu události, přičemž jednotkou času je měsíc. Třetí sloupec udává, zda ke studované události došlo, nebo bylo pozorování cenzorováno. R1 je z generace 1972, délka vystavení riziku výskytu události je u ní 108 měsíců od výchozí události (tedy od 15. narozenin) a sledovaná událost u ní

Tabulka 1. Příklad s vysvětlující proměnnou konstantní v čase

ID	TIME	EVENT	Rok narození
1	108	1	1972
2	180	0	1980
3	...	...	...

Tabulka 2. Příklad s vysvětlující proměnnou konstantní a měnící se v čase – episode splitting

ID	Start	End	TIME	EVENT	Vzdělání	Rok narození
1	0	42	42	0	ZŠ	1972
1	42	108	108	1	SŠ	1972
2	0	36	36	0	ZŠ	1980
2	36	60	60	0	SŠ	1980
2	60	120	120	0	VOŠ	1980
2	120	180	180	0	VŠ	1980
3	...	...	...	...	...	...

Poznámka: Ukázka úpravy datové matice v programu STATA.

nastala. V praxi to tedy znamená, že ve věku 24 let (108 měsíců po 15. narozeninách) se jí narodilo první dítě. R2 je z generace 1980, vystavena riziku události je po dobu 180 měsíců a událost u ní nenastala. V praxi tedy bylo respondentce v době rozhovoru 30 let (180 měsíců po 15. narozeninách) a byla bezdětná.

Tato jednoduchá datová matice se promění, zahrneme-li vysvětlující proměnnou vzdělání (tabulka 2). Pro analytický model rozdělíme jedno pozorování do více řádků tak, aby byla vždy jasně odlišena doba, po kterou byla respondentka vystavena riziku výskytu události a zároveň měla dosaženou danou úroveň vzdělání. Začátek a konec této epizody označují sloupce „Start" a „End", přičemž tento interval je zleva otevřený. U R1 takto upravený formát dat zachycuje, že prvních 42 měsíců (do věku 18,5 let) byla vystavena riziku události jakožto osoba s ukončeným základním vzděláním a po tuto dobu se jí dítě nenarodilo (tato epizoda pozorování je cenzorovaná, neboť v ní ke sledované události nedošlo). Ve věku 18,5 let dokončila střední školu a od 42. měsíce pozorování do 108. měsíce tedy byla vystavena riziku výskytu události již jako osoba se středoškolským vzděláním. Analogicky lze v datech odlišit měnící se úroveň vzdělání u respondentky 2.

Z uvedených základních pojmů vztahujících se k analytickým postupům a principů sledování jedince a jeho životních událostí plynou také jisté specifické požadavky na obsah a formát datových zdrojů, které mohou být k daným analýzám využity.

Pro analýzu historie událostí v sociálních vědách jsou nejčastěji využívána data z výběrových šetření (retrospektivních i prospektivních). Jedná se o tzv. **life/event history data**, která musí zahrnovat informace o studovaných udá-

lostech (např. pro studium zakládání rodiny jsou relevantní informace o roku či věku při dokončení studia, nástupu do zaměstnání, seznámení se s partnerem, počátku kohabitace, sňatku, narození prvního dítěte atd.) a pokud možno co nejpřesnější datace jednotlivých událostí. Při použití běžných časových jednotek (měsíc, rok, v některých případech i dny či hodiny) musí být měření času u všech studovaných jedinců shodné.

Výhodou použití dat z **retrospektivních šetření** je nepochybně menší náročnost jejich sběru (finanční i časová) a také pokrytí delšího časového období. Naopak nevýhoda může pramenit právě z kvality časových údajů. Retrospektivní šetření klade vyšší nároky na paměť respondentů, kteří mohou být dotazováni často i na události, které se staly před několika desítkami let. Datace získané z retrospektivních šetření mohou být tedy ovlivněny špatnou pamětí respondentů.[7] Je nutné počítat také s jistým limitem tolerance vůči objemu požadovaných údajů. V retrospektivních šetřeních lze zjišťovat především důležité životní události, nikoli však otázky motivační, postojové, kognitivní, zaměřené na afektivní stavy či volní faktory. Dále je nutné počítat s tím, že do výzkumného vzorku retrospektivních šetření se dostanou pouze „přežívající", nikoli však ti ze sledovaných generací, kteří již zemřeli či se např. odstěhovali.

Nespornou výhodou **panelových výzkumů**[8] jsou nižší nároky na paměť respondentů a šance získat přesnější datace, jsou-li události sledovány průběžně nebo s malým časovým odstupem. Zároveň je možné ptát se průběžně na plány, názory či motivaci respondentů. Nevýhodou prospektivních výzkumů jsou naopak enormní požadavky časové i finanční, neboť pro získání informací například o rodinném chování je nezbytné sledovat výzkumný vzorek často i několik desetiletí. S panelovými výzkumy jsou spojeny také statistické problémy spojené s nemožností udržet původní soubor dotazovaných kompletní a s efektem selekce, neboť po několika opakováních výzkumu již nemusí být původní reprezentativita souboru zaručena kvůli lidem, kteří z něj „vypadli". Diskutován je v této souvislosti například také panelový efekt a otázka, zda může opětovné dotazování ovlivnit odpovědi, příp. chování respondentů.

4. Základní statistické metody analýzy historie událostí

V analýze zaměřené na události je možné rozlišovat tři hlavní typy přístupů: neparametrické metody, parametrické a semi-parametrické metody.

Neparametrické deskriptivní metody

Metody označované jako **neparametrické** nepředpokládají apriori rozdělení studovaného jevu (studovaných jevů) a tudíž ani žádnou konkrétní funkci (danou

7 Kontrolní zkouška v Belgii, porovnávající data výběrového šetření s daty populačního registru, prokázala, že chyby v datování událostí byly časté, i když časový sled událostí byl správný. Bylo však ověřeno, že vliv chybných údajů na parametrickou i neparametrickou analýzu není podstatný, je-li zachován správný sled událostí [Lelièvre 1992].

8 K panelovým výzkumům více Monotematické číslo Panelový výzkum – Data a výzkum 2009, roč. 3, č. 1.

parametry) popisující daný jev. Tato metoda analýzy je srovnatelná s technikami tabulek života a dvě základní neparametrické metody jsou metoda Kaplan-Meier (*Product Limit Estimator*) a tabulky života (*Life Table Estimation*).

Představu o průběhu studovaného jevu si lze udělat na základě výpočtu **funkce přežití** (*survival function*) udávající podíl těch, u kterých do určitého časového okamžiku (na konci intervalu *t*) nenastala studovaná událost, resp. **stále zůstávají v riziku podstoupit studovanou událost**:

$$S(t) = P(T \geq t) = 1 - P(T < t),$$

kde T je datum podstoupení studované události, $P(T < t)$ je pravděpodobnost podstoupit studovanou událost před datem t.[9]

Pro popis funkce přežití se používá jak mediánů časů přežití, tak také podílů osob, které v daném čase ještě stále zůstávají v riziku podstoupit studovanou událost. Funkce přežití umožňuje porovnávat také odlišnosti mezi jednotlivými skupinami (např. pohlaví, generace, země) a testovat významnost pozorovaných rozdílů (např. neparametrickým log rank testem[10]). Při porovnávání mnoha kategorií je však nutné konstruovat funkce přežití pro mnoho různých podskupin, což v důsledku znemožňuje interpretaci výsledků. Funkce neparametrických metod je tedy především popisná v úvodu analýzy, chceme-li analyzovat vliv více vysvětlujících proměnných, je nutné konstruovat modely řadící se do skupin parametrických či semi-parametrických metod. V popisné analýze je grafické znázornění výsledných funkcí prvotní, zároveň je však něčím víc než pouhou vizualizací, neboť zjištění průběhu určitého jevu umožňuje volit další postup.

U **metody tabulek života** (*Life table, někdy též aktuárský odhad*) je funkce přežití rozdělena do určitého počtu intervalů za předpokladu, že ke studovaným událostem dochází na intervalu rovnoměrně. Tato metoda je vhodná spíše pro větší soubory, aby bylo zajištěno, že se ve zvolených intervalech budou vyskytovat studované události. Je vhodná v případě, že nejsou k dispozici dostatečně malé časové jednotky (např. máme-li pozorování pouze v letech, což více odpovídá charakteru intervalů), případně jsou známy pouze intervaly, ve kterých nastala událost nebo došlo k cenzorování pozorování. Výhodou je, že aktuárský odhad funkce přežití poskytuje informace o změnách rizika ve stejně dlouhých po sobě následujících intervalech pozorování, a mnohdy jeho větší jednoduchost.

Funkce přežití metodou tabulek života je počítána podle vzorce:

$$S(t) = \prod_{i/t(i) \leq t} [1 - d_i / (N_i - \tfrac{1}{2} m_i)],$$

kde d_i je počet studovaných událostí během intervalu t, N_i je počet respondentů na začátku intervalu, kteří mohou v čase t podstoupit studovanou událost, m_i počet cenzorovaných pozorování během intervalu. Odhad předpokládá, že v pří-

9 Symbolika používaná k popisu vybraných analytických funkcí je převzatá z Lelièvre, Bringé [1998].

10 Log rank test, též Coxův-Mantelův test, spočívá v porovnávání pozorovaných a odhadnutých počtů událostí ve všech okamžicích, kdy došlo alespoň k jedné události. Výsledné testovací kritérium je porovnáno s kritickou mezí χ^2-rozdělení [více viz Hendl 2004: 451–454].

padě cenzorovaného pozorování zůstávají jedinci vystaveni riziku studované události v průměru polovinu doby ze zvoleného časového intervalu.

Zásadní nevýhodou apriorní konstrukce intervalů je však to, že výsledné odhady závisejí na velikosti zvolených intervalů. Při moderní výpočetní technice již tedy není důvodem preferovat tuto metodu z důvodu úspory času či kapacity počítače [Blossfeld, Rohwer 2002].

V současné době se častěji používá odhad funkce přežití pomocí **metody Kaplan-Meier** (*Product Limit Estimator*). V metodě Kaplan-Meier je čas rozdělen do intervalů, z nichž každý obsahuje pouze jeden časový okamžik, ve kterém došlo ke studované události. Funkce přežití je tedy odhadnuta v každém okamžiku, kdy byla v souboru pozorována událost (ke studované události dochází v přesných časech $t_1, t_2, ..., t_n$). Časové okamžiky jsou tedy zachyceny každý individuálně. Výhodou této metody je, že hledané funkce nejsou závislé na způsobu definování časové proměnné a díky aktualizaci odhadu pokaždé, když dojde ke studované události, poskytuje nejpřesnější odhad. Funkce přežití je zde počítána podle vzorce:

$$S(t) = \prod_{i/t(i)\leq t} (1 - d_i / N_i),$$

kde $S(t)$ je odhadovaná funkce přežití, d_i je počet studovaných událostí v čase t, N_i je počet respondentů, kteří mohou v čase t podstoupit studovanou událost.

Regresní modely pro life history data – modely rizika

Uvedené neparametrické metody tedy poskytují informaci o časování daného jevu a odlišnostech při porovnání dvou či více podskupin. Cílem analýz je však většinou zahrnutí více proměnných a studium jejich vlivu na výskyt dané události. Proto byly vytvořeny **regresní modely**, které jsou založeny na myšlence „vysvětlit riziko podstoupení studované události v daném časovém okamžiku (vysvětlovaná proměnná) pomocí určitých charakteristik (vysvětlující proměnné)“ [Rychtaříková 2008: 254].

Parametrické a semi-parametrické metody umožňují analyzovat vliv řady nezávislých proměnných. Jedná se tedy o rozšíření standardních regresních modelů na analýzu délky trvání (*duration*), při níž je zároveň možné zohledňovat různorodost sledované populace. Přístupy založené na regresních modelech navíc umožňují studovat interakce mezi procesy a proměnnými i analyzovat skrytou heterogenitu (*unobserved heterogenity*) souboru.

Závislou proměnnou v těchto modelech je funkce rizika **h(t)** (*hazard rate function, také známá jako hazard function, risk, intensity of the event`s occurence, česky se používají též termíny riziková funkce, riziko*),[11] která udává profil, jak se v čase mění riziko, že dojde ke sledované události. Jedná se o míru

11 V textu uvažujeme pouze skupinu modelů, kde je závislou proměnnou funkce rizika, neboť tato závislá proměnná je nejvhodnější pro daný typ modelů. Existují také modely, které se soustředí na analýzu času trvání, ve kterých je závislou proměnnou délka trvání (survival time – accelerated failure-time modely) nebo funkce přežití [viz Vermunt 1997a]. Tyto nejsou v textu diskutovány.

rizika podstoupit studovanou událost ve velmi krátkém časovém intervalu délky Δt (mezi časem *t* a *t* + Δ*t*) za předpokladu, že k události nedošlo do okamžiku *t*.

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t / T \ge t)}{\Delta t},$$

kde *T* je datum podstoupení studované události, *f(t)* je hustota pravděpodobnosti (hustota rozdělení časů přežití), *S(t)* je funkce přežití. Funkce rizika v modelech se spojitým časem tedy není pravděpodobnost, je to míra (*rate*) nebo též okamžiková intenzita vztahující podmíněnou pravděpodobnost výskytu události ku velmi krátkému časovému intervalu.

Parametrické a semi-parametrické metody se od sebe liší tím, jaké předpoklady stanovují ohledně rozdělení funkce rizika v závislosti na čase. Je-li předpokládáno určité statistické rozdělení rizikové funkce, jedná se o parametrické modely, pokud není funkce rizika parametrizována (není formulován žádný předpoklad o rozdělení rizikové funkce), jedná se o semi-parametrické metody.

Semi-parametrické metody byly navrženy Coxem, často jsou tedy označovány jako Coxův semi-parametrický model. Cox představil původní model v roce 1972 a spojil v něm regresi s tabulkami života a tento model je považován za „zavedení dynamiky do regrese, nebo alternativně, jako metoda pro měření dopadu proměnných v procesu transformace a analýzy rizik" [Lelièvre, Bringé 1998: 109].

Coxova regrese odhaduje rizikovou funkci dle následujícího vztahu:

$$h(t, x) = h_0(t) \exp \beta x,$$

kde *h(t;x)* je funkce rizika v okamžiku *t* pro jedince s charakteristikami *x*, $h_o(t)$ je základní (též bázová) funkce rizika (baseline), jsou-li všechny vysvětlující proměnné 0 (*x* = *0*), *x* jsou vysvětlující proměnné a *β* příslušné regresní koeficienty pro vliv proměnných na intenzitu procesu. Jelikož model nepracuje s předpoklady ohledně rozdělení bázové funkce, která není blíže specifikována, ale stanovuje předpoklady, jak sledované nezávislé proměnné působí na rizikovou funkci, je označován jako semi-parametrický. Bázová funkce rizika se tak v rámci Coxovy regresní analýzy nemusí odhadovat a regresní koeficienty *β* jsou odhadovány nezávisle na funkci rizika (metodou maximalizace parciální věrohodnosti, která je založena pouze na pořadí sledovaných událostí).

Coxův model však klade předpoklady toho, jak sledované nezávisle proměnné působí na rizikovou funkci – vychází z předpokladu proporcionálního vlivu, a řadí se tak do kategorie modelů proporcionálních rizik, protože vliv vysvětlujících proměnných může přivodit proporcionální posuny funkce rizika (funkce rizika se pouze posune směrem k vyšším nebo nižším hodnotám [Hendl 2004:456]), nemohou však změnit její tvar [Blossfeld, Rohwer 2002]. Pro konkrétní aplikaci Coxova modelu je tedy nutné předpoklad proporcionality ověřit.[12]

12 Jedním ze způsobů je grafická metoda založená na transformaci odhadů funkce přežití pomocí funkce log(–log*S(t)*). Je-li předpoklad proporcionality splněn, křivky by měly být paralelní a jejich rozdíl by měl být konstantní v průběhu sledovaného času.

Právě toto tvoří jedno ze dvou slabých míst semi-parametrických modelů, neboť předpoklad proporcionality, na kterém je založen odhad regresních koeficientů při neznámém rozdělení rizikové funkce, je v řadě aplikací tohoto modelu nereálný. Druhou slabinou je pak chybějící odhad základní funkce rizika, neboť v řadě analýz je důležitým výstupem právě průběh funkce rizika v závislosti na čase (zajímá nás, jak se v průběhu času mění riziko výskytu studované události). Řešením těchto dvou problémů však může být zahrnutí interakce času s dalšími proměnnými v modelu nebo zahrnutí času trvání sledovaného procesu v podobě v čase se měnící proměnné (time-varying covariate) [Vermunt 1997a].

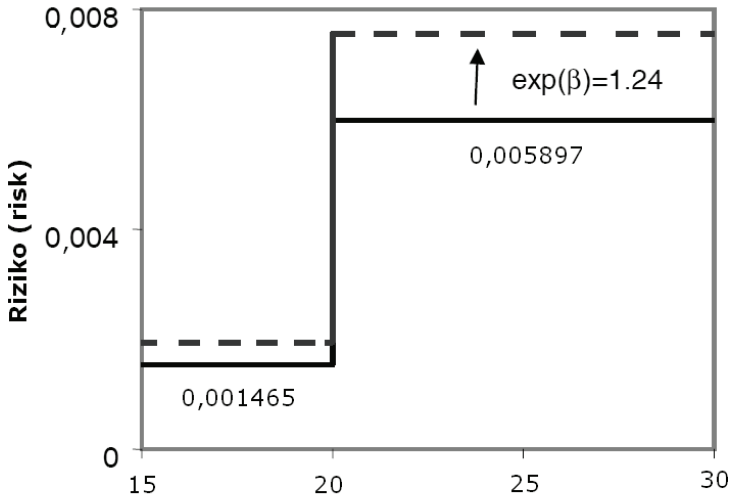
Parametrické metody představují skupinu modelů, které pracují se známým statistickým rozdělením (např. exponenciální, logistické, gamma) aplikovaným na sledovanou distribuci funkce přežití a funkce rizika. Některé modely pracují také s předpoklady ohledně vlivu individuálních charakteristik. Odhad parametrů těchto modelů probíhá metodou maximální věrohodnosti. V rámci této skupiny modelů jsou často používány modely exponenciální, Weibull, Gompertz-Makeham, log-logistic, log-normal atd., kterým však v tomto článku nebude věnována větší pozornost, neboť v sociálních vědách často není možné u studovaného jevu určit odpovídající statistické rozdělení, na kterém při použití parametrických modelů závisí výsledné odhadované parametry (více ke specifikaci jednotlivých parametrických modelů viz např. Hosmer, Lemeshow 1999, Lelièvre, Bringé 1998, Blossfeld, Rohwer 2002 nebo Vermunt 1997a).

Zaměříme se však na model, který je možné řadit do skupiny parametrických modelů (konkrétně exponenciálních modelů) a který se pohybuje na pomezí mezi jinými parametrickými modely, které pracují se striktními předpoklady ohledně rozdělení rizikové funkce, a Coxovou regresí, která naopak žádný takovýto předpoklad nemá. Tzv. **piecewise constant exponenciální model**, jehož výsledky budou prezentovány v následující části článku pro ilustraci interpretačních možností prezentovaných metod, je vhodným modelem využívaným při analýzách pracujících s časem jako spojitou proměnnou.

Tento model je jednoduchou generalizací exponenciálního modelu (v exponenciálním modelu je riziková funkce v čase konstantní, což příliš neodpovídá v realitě studovaným jevům), neboť riziková funkce je zde konstantní na zvolených intervalech, ale proměnlivá mezi těmito intervaly. Model je tedy velmi flexibilním nástrojem, především pokud nemá výzkumník jasnou představu o rozdělení sledovaného jevu, neboť při vhodné volbě intervalů umožňuje modelovat téměř každou základní (bázovou) funkci rizika. Navíc je možné volit krátké intervaly tam, kde riziko výskytu studovaného jevu výrazně variuje, a širší intervaly tam, kde se riziko mění pouze pozvolna (např. pokud se riziko propuštění ze zaměstnání v daném oboru výrazně mění, stoupá či klesá v prvním roce po nástupu do zaměstnání a v dalších letech již výrazněji nekolísá, je možné zvolit v prvním roce intervaly např. měsíční a poté volit roční i víceleté intervaly).

Piecewise constant exponenciální model, ve kterém mají vysvětlující proměnné stejný efekt po všechny sledované časové intervaly, patří taktéž k modelům

Obrázek 2. Ilustrace předpokladu proporcionality rizika – piecewise constant exponencial model



proporcionálních rizik[13] [Blossfeld, Rohwer 2002]. Pokud tedy zapíšeme model proporcionálního rizika obecně:

$$h(t, x) = h_0(t) \exp \beta x ,$$

pak při rozdělení délky trvání J body na časové ose $0 = \tau_0 < \tau_1 < \tau_2 ... < \tau_J = \infty$ je možné definovat j-tý interval jako $(\tau_{j-1}, \tau_j]$ a zapsat bázovou funkci rizika konstantní na každém intervalu [Rodríguez 2007]:

$$h_0(t) = h_j \text{ pro } t \text{ v intervalu } (\tau_{j-1}, \tau_j] .$$

Při volbě intervalů je nutné uvážit jejich množství – pokud zvolíme velké množství intervalů, dostaneme sice lepší odhad neznámé funkce rizika, ale nese to s sebou nutnost odhadu velkého počtu koeficientů, je-li naopak zvoleno málo intervalů, je sice méně problémů s odhadem, ovšem nese to s sebou riziko nedostatečné aproximace rizikové funkce. Ve většině případů je tedy nutné volit jistý kompromis s podmínkou, že v každém zvoleném intervalu by měly být nějaké pozorované události (tedy časy výskytu události) [Blossfeld, Rohwer 2002]. V teoretickém případě, kdy by se počet intervalů rovnal počtu diskrétních časů, ve kterých se vyskytla studovaná událost, piecewise constant model by se rovnal popsanému semi-parametrickému modelu [Vermunt 1997a]. V praxi při dobré specifikaci tohoto modelu dávají jeho výsledky obdobné odhady parametrů jako Coxův model, s tou výhodou, že v piecewise constant modelu je zároveň odhadnut průběh funkce rizika.

13 Proporcionalita rizik není specifickým rysem pouze Coxovy regrese, ač Coxův model nese mj. tento název. Proporcionální modely rizika (*proportional hazard models*) tvoří skupinu modelů, ve které jsou zastoupeny i některé parametrické modely [Blossfeld, Rohwer 2002; Vermunt 1997a].

Ilustraci proporcionality rizika je možné znázornit při zahrnutí v čase se neměnících proměnných. Na obrázku 2 je znázorněna část funkce rizika, která je konstantní na daném intervalu, mezi intervaly se však mění. Nepřerušovanou čarou je znázorněna funkce rizika pro referenční kategorii – baseline (např. muži), přerušovanou čarou je znázorněno, že riziková funkce srovnávané kategorie (žen) je násobkem rizikové funkce mužů jakožto referenční kategorie, riziko výskytu studované události u žen je proporční vzhledem k referenční kategorii a vyšší o 24 % v porovnání s muži.

Piecewise constant model je flexibilní také v tom, že umožňuje zahrnutí vysvětlujících proměnných měnících se v čase i interakce proměnných s rizikovou funkcí. V těchto případech je pak možné hovořit o neproporcionálních efektech nebo též o v čase proměnlivých efektech.

5. Příklady výstupů uvedených metod a jejich interpretací

Výstupy vybraných metod budou nyní demonstrovány na analýze procesu odchodu mladých dospělých z domácnosti rodičů. V souboru 9 029 českých mužů a žen (4 340 mužů a 4 689 žen) z generací 1926–1987[14] sledujeme odchod od rodičů jakožto jednu z událostí širšího procesu nazývaného přechod do období dospělosti. Tuto událost sledujeme od 15. narozenin respondentů, pozorování jsou cenzorována okamžikem šetření nebo ve věku 35 let. Časovou proměnnou je doba uplynulá od 15. narozenin počítaná v měsících. Jelikož nás zajímá vedle časování události také vliv vybraných proměnných, jsou do modelu zahrnuty následující vysvětlující proměnné: generace, charakteristika orientační rodiny v době dětství (typ rodiny, ve které respondent převážně žil do svých 15 let a počet sourozenců) a dále proměnné měnící se v čase: nejvyšší dokončené vzdělání a partnerský status. Pro bližší specifikaci souboru, proměnných, jejich operacionalizace jakož i celkové výsledky viz Kuchařová, Šťastná [2009].[15]

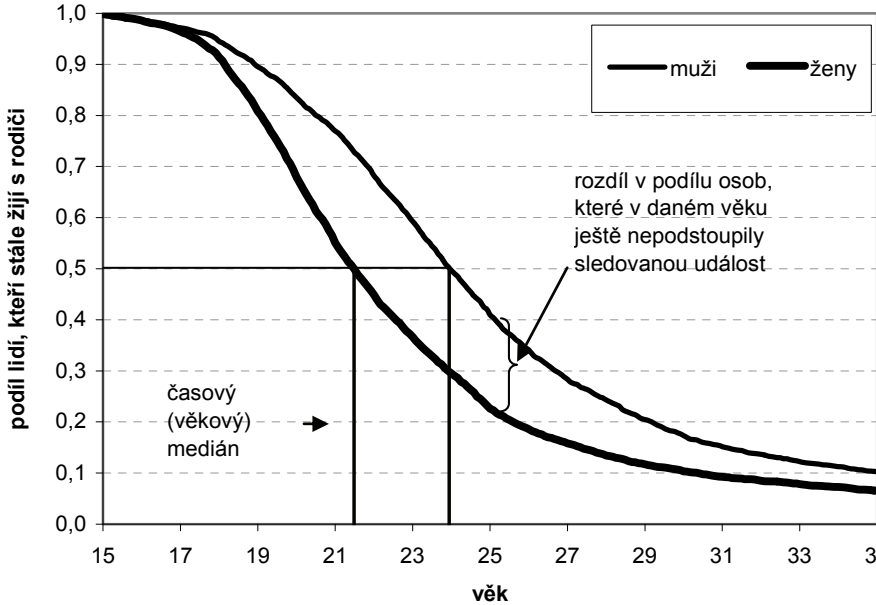
Na obrázku 3 je znázorněn výstup neparametrické metody studia historie událostí – funkce přežití udávající podíl těch, kteří v daném věku stále žijí ve společné domácnosti s rodiči. Křivky jsou vypočítány zvlášť pro muže a pro ženy a jsou porovnány graficky, jejich odlišnosti jsou testovány také statisticky. Pro popis dat se pak používá jak mediánů času přežití, tak srovnávání podílu osob, které v určitém okamžiku, případně na konci pozorování ještě nepodstoupily studovanou událost. Prezentované výsledky nám konkrétně ukazují rozdíl v mediánu času přežití (zde věku, ve kterém opustí domácnost rodičů polovina žen, resp. mužů). Ten je pro ženy 21,5 let, pro muže o dva a půl roku vyšší, tedy 24 let. Dále můžeme v každém věku porovnávat podíly osob, které doposud žijí ve společné domácnosti s rodiči, např. ve věku 25 let je to 23 % žen a 41 % mužů.

Jako model zahrnující vysvětlující proměnné byl použit konkrétně piecewise constant model popsaný v části 4 tohoto článku. Tento model byl volen proto,

14 Jsou použita data výběrového šetření Muži a ženy v ČR: životní dráhy a mezigenerační vztahy 2005.

15 Kompletní model odhadnutý autorkou textu je dostupný v Kuchařová, Šťastná [2009].

Obrázek 3. Odchod mužů a žen z domácnosti rodičů (funkce přežití)



Poznámka: Metoda Kaplan-Meier, Log Rank test rozdílnosti křivek signifikantní při $p < 0{,}001$.

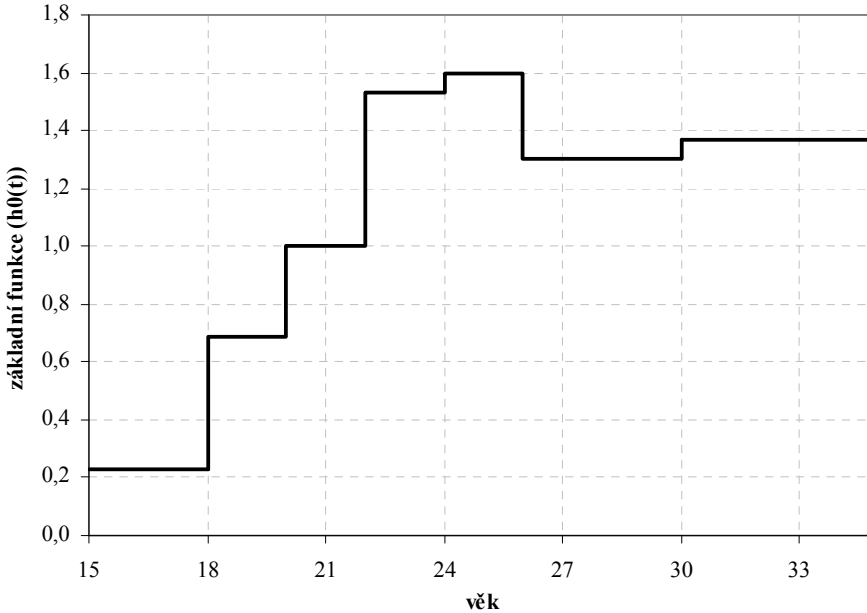
že předmětem studia bylo jak časování daného procesu (a bylo tedy nutné odhadnout průběh základní funkce), tak vliv vysvětlujících proměnných na tento proces, z nichž některé se mohly v průběhu sledovaného času měnit (konkrétně nejvyšší ukončené vzdělání, rodinný stav).

Na obrázku 4 (na následující ztraně) je znázorněna intenzita[16] odchodu mužů z domácnosti rodičů (tedy základní funkce – baseline), na které je z jednotlivých „skoků" patrné, do jakých věkových intervalů byla v tomto případě rozdělena sledovaná doba trvání procesu. Z výsledků je patrné, že nejvyšší intenzita odchodu mužů z domácnosti rodičů je mezi 24. a 26. rokem věku a dále ve věku 22 a 23 let. V mladším věku je intenzita výrazně nižší, klesá také po 26. narozeninách. Mírné zvýšení je patrné opět u třicetiletých a starších mužů.

Pro komentování dalších výsledků vybíráme jednu v čase konstantní a jednu v čase měnící se proměnnou. Vliv typu rodiny, resp. její úplnosti v době dětství

16 V interpretaci výsledků užíváme slovo „intenzita" alternativně ke slovu „riziko", neboť v českém překladu nezní riziko ve spojení s některými studovanými procesy nejvhodněji (např. „riziko sňatku", „riziko návratu do zaměstnání"). Zároveň, jak bylo uvedeno v předešlé části, riziková funkce je také v angličtině někdy označována jako „intensity of the event's occurence".

Obrázek 4. Intenzita odchodu mužů z domácnosti rodičů (Hazard risk)



respondenta se na osamostatňování mužů projevuje tak, že téměř dvojnásobně (o 93 %) zvyšuje riziko odchodu z domácnosti rodiče pro ty muže, kteří vyrůstali v rodině s jedním vlastním a jedním nevlastním rodičem, oproti respondentům z úplných rodin. Riziko odchodu pro muže z neúplných rodin není nijak signifikantně odlišné od osob z úplné rodiny (viz tabulku 3).

Proměnná vzdělání, která je v modelu zahrnuta jako v čase se měnící proměnná, naznačuje, že v průběhu studia se mladí lidé od svých rodičů stěhují nejméně často (riziko je o 38 % nižší oproti referenční kategorii nestudujících mužů s ukončeným středoškolským vzděláním s maturitou). S výší dosaženého vzdělání však intenzita opuštění domácnosti rodičů statisticky významně roste – vysokoškoláci vykazují 1,4krát vyšší intenzitu odchodu než středoškoláci s maturitou. Ačkoli tedy delší doba studia spojená se získáním vyššího vzdělání oddaluje odchod z domova rodičů, po ukončení procesu vzdělávání je osamostatnění se z domácnosti rodičů nejintenzivnější u vysokoškolsky vzdělaných mladých mužů. Muži s nižším vzděláním bez maturity se naopak z domácnosti rodičů stěhují s nižší intenzitou (jejich riziko odchodu od rodičů je o 20 % nižší než v případě mužů s maturitou - viz tabulku 3).

Zároveň je nutné zdůraznit, že vliv jednotlivých proměnných je očištěn od korelace s jinými vysvětlujícími proměnnými. Výsledky popisující efekt vzdělání tedy prezentují „čistý vliv“ vzdělání za předpokladu, že by se sledovaní respondenti nelišili v jiných charakteristikách zahrnutých (kontrolovaných) v modelu.

Tabulka 3. Intenzita odchodu od rodičů, vliv vybraných proměnných, muži

	exp(β)	signifikance
Vzdělání		
studující	0,62	***
Ukončené studium:		
základní	0,80	***
střední bez maturity	0,89	**
střední s maturitou	*1*	
vysokoškolské	1,39	***
neudáno	0,83	*
Domov rodičů v době dětství		
s oběma biologickými rodiči	*1*	
1 biologický rodič, 1 nevlastní	1,93	***
pouze 1 biologický rodič	1,09	
jinak/nezjištěno	1,07	

Poznámka: Metoda: event-history analýza (piecewise constant model). Signifikance: *** p < 0,01; ** p < 0,05; * p < 0,1.

Z ukázky je zřejmé, že regresní koeficienty β většinou nejsou interpretovány přímo, ale ve tvaru exp(β), a takto umocněný koeficient lze interpretovat jako **relativní riziko**. Zvolená referenční kategorie nabývá hodnoty 1, porovnávané kategorie mohou nabývat nižších či vyšších hodnot vzhledem k dané referenční. Relativní riziko (exp(β)) však nikdy nemůže být záporné.

6. Použití metod event history analýzy při studiu (české) populace

Doposud prezentované příklady byly čerpány především z oblasti populačních studií. Oblasti, ve kterých je možné v rámci sociálních věd zmíněné analytické metody využívat, se však neredukují pouze na analýzu výskytu důležitých demografických událostí. Již v úvodu bylo zmíněno, že uvedené metody se často pod jinými názvy aplikují v technických vědách při analýzách spolehlivosti, resp. poruchovosti (např. při hodnocení životnosti technických součástek), ve vědních oborech, jako je medicína a epidemiologie (např. délka přežívání při stanovené diagnóze, relaps onemocnění). V oblasti sociálních deviací může být sledován např. i relaps v případě drogových závislostí a jejich léčení. V rámci sociálních věd proto uveďme pro ilustraci další oblasti a procesy, které je vhodné zkoumat za využití analýzy historie událostí.

Již bylo naznačeno, že v sociologii se otevírá široké pole možností analýz v oblasti sociologie rodiny v návaznosti na demografické procesy. Analytické postupy lze však aplikovat také např. při studiu sociální mobility i na témata přesahující do sféry sociální politiky, jako jsou otázky změny životní úrovně (např. s přechodem do důchodu či v seniorském věku), výstup nad hranici chudoby či naopak sestup pod ni apod. Obdobné využití je možné při studiu prostorové mobility obyvatelstva (migrace vnitřní i zahraniční).

V psychologii se uvedené metodické postupy mohou využívat ke studiu vývojového procesu, v psychiatrii například ke studiu lidí vykazujících znaky psychózy nebo neuróz a případně propuknutí nebo relapsu psychiatrického onemocnění. V oblasti pedagogického výzkumu a vzdělávání lze sledovat procesy, jakými je (ne)ukončení studované školy, volba různých vzdělávacích strategií (např. přechody mezi typy škol a úrovní vzdělání), celoživotního vzdělávání.

Široké pole působnosti se otevírá na poli ekonomie i aplikované ekonomie (např. podnikové ekonomie), kdy může předmětem zkoumání být časování návratu do zaměstnání, pohyby mezi zaměstnaností a nezaměstnaností, mezi částečnými a plnými úvazky, ale i setrvání zaměstnanců v konkrétním podniku či na konkrétní zaměstnanecké pozici. V marketingovém výzkumu lze sledovat například odklon konzumentů k jiné značce. Metody mají uplatnění například také v oblasti kriminologie a vězeňství, např. při studiu recidivy (sledována je doba mezi propuštěním do návratu do vězení/spáchání dalšího deliktu).

V posledním desetiletí se s aplikací event history metod setkáváme stále častěji také u českých autorů, kteří pro analýzu socio-demografických témat využívají české datové soubory. Zmiňme se proto na závěr o tématech, která jsou v našem prostředí touto metodou studována. Cílem není podat úplný výčet autorů a jejich prací jako spíše upozornit na významné tuzemské datové zdroje a hlavní témata, jimž je věnována pozornost.

Jedním z témat je **přechod do dospělosti** a sledovány jsou proměny zejména v kontextu měnících se ekonomických a společenských podmínek České republiky v posledních 20 letech. Skutečnost, že vzdělávací, pracovní a rodinné dráhy mladých dospělých po roce 1990 nesledovaly výrazně uniformní dráhu, jako tomu bylo u Češek, které vstupovaly do dospělosti v 70. a 80. letech, využila Kantorová [2004] jako rámec své analýzy formování prvního partnerského stavu a narození prvního dítěte. Explicitní pozornost přitom věnovala roli vzdělání žen a jejich zaměstnanosti při porovnávání dvou historických období: období 70. a 80. let 20. století a období socio-ekonomické transformace v 90. letech.

Hamplová [2003] analyzovala vztah mezi formováním partnerského soužití a vzděláním žen, Škop [2005] pak proces osamostatňování se mladých lidí a jejich odchod z domácnosti rodičů. Všichni tito autoři analyzovali datový soubor Šetření rodiny a reprodukce (FFS 1997), který obsahuje údaje o 1 735 ženách narozených v letech 1952 a 1982. Paloncyová [2002] aplikovala metody event history analýzy na data sbíraná v rámci vlastního výzkumného projektu „Biografický výzkum mladé generace 2002“ zaměřeného na životní dráhy mladých lidí narozených v letech 1968 a 1977. Zaměřovala se přitom především na sledování zahájení partnerského soužití a manželství, oblast vzdělání, vstup na trh práce a profesní kariéry.

Druhým z témat je **plodnost** a krom již uvedených studií zabývajících se začátkem rodičovské fáze, tedy narozením 1. dítěte a vstupem do rodičovství [Kantorová 2004, Paloncyová 2002], je sledováno také rození dětí vyššího pořadí. Pikálková [2003] se soustředila na analýzu podmínek a kontextu narození třetího dítěte, zejména na vztah mezi narozením třetího dítěte a úrovní vzdělání

matky. Také ona používala k analýzám data Šetření rodiny a reprodukce (1997). Na sledování souvislostí a determinant narození dětí druhého pořadí se zaměřuje ve svých analýzách Šťastná [2009], která využívá data z novějšího výzkumu Muži a ženy v ČR: životní dráhy a mezigenerační vztahy (GGS 2005).

Třetí oblastí objevující se v pracích českých autorů je problematika **rozvodovosti**. Zeman [2003] se zaměřil na analýzu rozvodovosti a rozpadu manželství nejen v České republice, ale také v Rakousku. Hlavní důraz klade na otázku předmanželského soužití a jeho roli v následné (ne)stabilitě manželství. Rozpad manželství je také studován s ohledem na kontext předešlé životní dráhy jedince a zejména v souvislosti s procesem odchodu z domácnosti rodičů a formováním partnerství. Šťastná [2005, 2006] se v rámci daného tématu zaměřila na otázku mezigeneračního přenosu rozvodového chování. Fučík [2007] propojuje oblast rozvodovosti a plodnosti, když studuje reprodukční strategie žen po rozvodu.

Dalšími aplikacemi event history analýzy je oblast **zaměstnanosti a pracovního** trhu se zaměřením na modelování délky nezaměstnanosti, které čerpají data z Výběrového šetření pracovních sil [Jarošová 2006, Malá 2007].

7. Doporučená literatura a software

Moduly k analýze historie událostí jsou součástí všech rozšířených statistických softwarů. Analýzu přežívání je možné počítat nejen v programu SAS, Stata a SPSS, ale také v programech jako je R, TDA (Transition Data Analysis[17]), aML[18] či lEM[19], které jsou mnohdy volně dostupné, jejich použití je však často obtížnější vzhledem např. ke specifickým požadavkům na formát dat, psaní programu a jeho spouštění (např. v aML). Uživatelsky pohodlnější a zároveň v sociálních vědách v ČR nejčastěji používané softwary (SPSS, SAS a Stata) nabízejí jak možnost volby příkazů z menu v dialogových oknech, tak programové okno pro psaní a spouštění napsaných syntaxí (např. syntax file v SPSS, log file ve Stata). Všechny obsahují také základní neparametrické metody (odhady funkcí přežití dle metody tabulek života i Kaplan-Meier). Popis dostupných semi-parametrických a parametrických metod a výklad konkrétních příkazů je k nalezení v on-line dokumentacích ke všem třem programům a v integrovaných nápovědách.

Stručným úvodem do metod analýzy historie událostí, dostupným v češtině, je kapitola v učebnici statistiky od J. Hendla [2004] či překlad přehledového textu od francouzské autorky dlouhodobě se zabývající metodami event-history analýzy É. Lelièvre [1992]. S dalšími zmínkami v českém jazyce je možné se setkat v textech českých autorů, kteří pomocí zmiňovaných metod analyzují vybrané jevy (viz část 6 textu), ovšem tyto zmínky jsou vhodné pro pochopení

17 Program napsaný G. Rohwerem a U. Pötterem v roce 2000 je volně dostupný na internetu ke stažení (http://www.stat.ruhr-uni-bochum.de/tda.html).

18 Program určený pro víceúrovňové modelování (multilevel and multiprocess models) napsaný autory Lillard, Panis [2003]. Volně dostupný na http://www.applied-ml.com/

19 lEM = log-linear and event history analysis with missing data using the EM algorithm [Vermunt 1997b].

analyzovaného procesu, nikoli jako učební text. Pro studium se musí zájemci obrátit k cizojazyčným textům, z nichž některé představují výklad určený i pro začátečníky v dané technice. Podrobným textem je kniha Judith D. Singer a Johna B. Willetta [2003] *Applied Longitudinal Data Analysis: Modelling Change and Event Occurrence,* která čtenáře seznamuje vedle metod analýzy historie událostí také s analýzou longitudinálních dat víceúrovňovými modely.

Mezi základní přehledové texty, které napomohou orientaci začátečníků v dané problematice, patří dále například Allison [1984] či manuál od autorů Lelièvre, Bringé [1998], kde jsou vysvětleny základní požadavky na data, postup při výpočtu ve třech softwarech (SAS, TDA a STATA) i interpretace výsledků. Základní postupy v uvedených programech jsou stále platné, bez ohledu na dataci citovaného manuálu. Každý softwarový produkt však v současné době nabízí podrobnou učebnici metod studia historie událostí, včetně relevantních příkazů a softwarových možností. Tyto učebnice bývají doplněny také dostupnými datovými soubory, které je možné využít pro vlastní zkoušení příkladů analyzovaných v učebních textech.

Software SAS nabízí například samostatnou knihu Paula Allisona [2010] *Survival Analysis Using SAS: A Practical Guide*. Software STATA nabízí hned několik textů z poslední doby. Jedním z nich je *Event History Analysis with Stata* od autorského kolektivu Blossfeld, Golsch a Rohwer [2007]. Nejnovějším textem je *An Introduction to Survival Analysis Using Stata* od autorského kolektivu Cleves, Gould, Gutierrez a Marchenko [2010]. Dalším zdrojem mohou být studijní materiály prof. Stephena P. Jenkinse [2008] k výukovému kurzu analýzy přežití s využitím programu Stata, jejichž součástí je taktéž možnost stáhnout si dostupné datové soubory a vyzkoušet modelové příklady.

SPSS má témata analýzy zpracovaná v rámci uživatelských příruček věnovaných pokročilým statistickým metodám, např. Norušis [2010], kde jsou kapitoly věnovány oběma metodám odhadu křivek přežití (Kaplan-Meier a tabulky života) i Coxově regresi.

Mezi klasické texty k event-history analýze, které však mohou být pro čtenáře náročnější, patří především Yamaguchi, K. [1991], Hosmer, Lemeshow [1999], Blossfeld, Rohwer [2002] či Courgeau, Lelièvre [1997].

8. Závěr

Studium biografií a vztahů mezi životními událostmi zůstávalo dlouhou dobu doménou kvalitativních sociologických studií, které disponovaly podrobnými životními historiemi malého počtu jedinců. Snahy analyzovat životní události v jejich vzájemných souvislostech postupem času vyústily ve vytvoření nových technik, které jsou založeny na specifických souborech kvantitativních dat a na odpovídajících statistických metodách. Pro takovéto analýzy jsou nezbytná data pokrývající historii životních událostí, která poskytují potřebné chronologické informace včetně časování jednotlivých událostí. Jedním z cílů tohoto článku bylo podat informaci o datových možnostech a analyzovaných tématech v českém prostředí. Existuje několik retrospektivních výzkumů, které shromáždily

údaje o životní dráze lidí, a také v současné době probíhají výzkumy, v nichž je kladen důraz na longitudinální přístup a které mají za cíl studovat dynamiku vývoje rodiny a rodinných vztahů od jejího založení až do jejího zániku.

Datové soubory (jak retrospektivního, tak panelového charakteru), na jejichž základě je možné zkoumat otázky související s dynamikou sociálních jevů a kde je možné aplikovat přístupy analýzy historie událostí, tedy v našem prostředí existují a v budoucnu budou přibývat. Je proto vhodné věnovat celému konceptu životní dráhy a jejímu výzkumu pozornost. Ať již kvůli explanačnímu potenciálu zmíněného přístupu, konceptuálním otázkám empirického zkoumání, či kvůli otázkám studia a rozvoje analytických metod.

Literatura

Alan, J. 1989. *Etapy života očima sociologie*. Praha, Panorama.

Allison, P. 1982. „Discrete-time methods for the analysis of event histories." *Sociological Methodology* 13: 61–98.

Allison, P. 1984. *Event History Analysis. Regression for Longitudinal Event Data*. London, New Delhi: SAGE Publications.

Allison, P. 2010. *Survival Analysis Using SAS: A Practical Guide*. Second Edition. SAS Press ISBN: 978-1-59994-640-5.

Blossfeld, H.-P., G. Rohwer. 2002. *Techniques of Event History Modeling. A New Approaches to Causal Analysis*. London: Lawrence Erlbaum Associates Publishers.

Blossfeld, H.-P., K. Golsch a G. Rohwer. 2007. *Event History Analysis with STATA*. Routledge Academic.

Cleves, M., W. Gould, R. G. Gutierrez, Y. V. Marchenko. 2010. An Introduction to Survival Analysis Using Stata. 3rd Edition, Stata Press.

Courgeau, D., É. Lelièvre. 1997. Event history analysis in demography. Oxford: Clarendon Press.

Fučík, P. 2007. „Porody po rozvodu". Pp. 171–184 In *Sociální reprodukce a integrace: ideály a meze*. Brno: Masarykova univerzita.

Hamplová, D. 2003. „Marriage and Educational Attainment: A Dynamic Approach to First Union Formation." *Sociologický časopis/Czech Sociological Review* 39(6): 841–863.

Hendl, J. 2004. *Přehled statistických metod pracování dat. Analýza a metaanalýza dat*. Praha: Portál.

Hoem, J. M., M. Kreyenfeld. 2006. „Anticipatory analysis and its alternatives in life-course research. Part 1: Education and first childbearing." *Max Planck Institute for Demographic Research Working Paper 2006-006*. Rostock: Max Planck Institute for Demographic Research.

Hosmer, D. W., S. Lemeshow. 1999. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. John Wiley & Sons.

Chaloupková, J. 2009. *Rodinné a pracovní dráhy mladých: holistická perspektiva*. Sociologické studie/Sociological Studies 09:07. Praha: Sociologický ústav AV ČR.

Chaloupková, J. 2010. „Výzkum životní dráhy a analýza sekvencí: možnosti studia životních drah." *Data a výzkum* 3(2): 241–258.

Jarošová, E. 2006. „Modelování délky trvání nezaměstnanosti“. *Statistika* 86(3): 240–251.

Jenkins, S. P. 2008. *Survival Analysis with Stata*. University of Essen, Institute for Social and Economic Research. Dostupné na http://www.iser.essex.ac.uk/study/resources/module-ec968.

Kantorová, V. 2004. *Family life transitions of young women in a changing society: First union formation and birth of first child in the Czech Republic, 1970-1997*. Doctoral thesis, Charles University in Prague and Universite de Paris I – Pantheon – Sorbonne. Available at: http://www.demogr.mpg.de/publications/files/1785_1000000000_1_Full%20Text.pdf .

Kuchařová, V., A. Šťastná (eds.). 2009. *Partnerství, rodina a mezigenerační vztahy v české společnosti*. Praha: PřF UK a VÚPSV.

Lelièvre, É. 1992. „Nové metody studia vztahů mezi demografickými událostmi.“ Pp. 225–237 in: Pavlík, Z. (Ed.). *Sňatečnost a rodina*. Praha: Academia.

Lelièvre, É, A. Bringé 1998. *Practical Guide to Event History Analysis using SAS, TDA, STATA*. Paris: INED.

Lillard, L. A., C. W. A. Panis. 2003. *aML Multilevel Multiprocess Statistical Software, Version 2.0*. EconWare, Los Angeles, California.

Malá, I. 2007. „Neparametrický odhad rozdělení doby nezaměstnanosti“. *Acta Oeconomica Pragensia* 2007, 15 (1): 52–62.

Manting, D. 1994. *Dynamics in marriage and cohabitation. An inter-temporal, life course analysis of first union formation and dissolution*. Amsterdam: Thesis publishers.

Norušis, M. J. 2010. *PASW Statistics 18 Advanced Statistical Procedures*. Pearson.

Pakosta, P., P. Fučík. 2009. „Vybrané metody analýzy panelových dat“. *Data a výzkum – SDA Info* 3(1): 77–96.

Paloncyová, J. 2002. *Rodinné chování mladé generace. Závěrečná zpráva z Biografického výzkumu mladé generace 2002*. Praha: VÚPSV.

Pikálková, S. 2003. „A Third Child in the Family: Plans and Reality among Women with Various Levels of Education.“ *Sociologický časopis/Czech Sociological Review* 39(6): 865–884.

Rodríguez, G. 2007. Survival Models. Kapitola ke kurzu Generalized Linear Models, Princeton Univerzity. Available at: http://data.princeton.edu/wws509/notes/.

Rychtaříková, J. 2008. „Nové metody demografické analýzy.“ *Demografie* 50(4): 250–258.

Singer, J. D., J. B. Willett. 2003. *Applied Longitudinal Data Analysis. Modeling Change and Event Occurence*. Oxford University Press.

Škop, M. 2005. *Statistická analýza přežívání s aplikací na proces odchodu od rodičů v České republice*. Doctoral thesis, Charles University in Prague.

Šťastná, A. 2005. „Mezigenerační přenos rozvodového chování na příkladu České republiky a v mezinárodním srovnání“. *Demografie*, 47 (1): 21–31.

Šťastná, A. 2006. „Rozvody a děti – vliv rozvodu rodičů na životní dráhu dětí.“ Pp. 175–190 in: Hamplová D., P. Šalamounová, G. Šamanová (eds.): *Životní cyklus – sociologické a demografické perspektivy*. Praha: Sociologický ústav AV ČR.

Šťastná, A. 2009. „Second Births in the Czech Republic.“ *Romanian Journal of Population Research* 3(1): 109–130.

Šubrt, J. 1993. „K vývoji názorů na problém času v sociologii.“ *Sociologický časopis*, 29(4): 471–492.

Vermunt, J. K. 1997a. *Log-linear Models for Event Histories*. *Sage Publications*.

Vermunt, J. K. 1997b. *LEM: A General Program for the Analysis of Categorical Data*. Department of Metodology and Statistics, Tilburg University.

Willekens, F. J. 1999. „The Life Course: Models and Analysis.“ in van Wissen, L. J. G., P. A. Dykstra (eds.). *Population Issues. An Interdisciplinary Focus*. New York: Plenum Publishers.

Yamaguchi, K. 1991. *Event History Analysis*. London: Sage Publications.

Zeman, K. 2003. *Divorce and marital dissolution in the Czech Republic and Austria. The role of premarital cohabitation*. Doctoral thesis, Charles University in Prague.