# POUŽITÍ KONEČNÝCH SMĚSÍ PRAVDĚPODOBNOSTNÍCH ROZDĚLENÍ PRO MODELOVÁNÍ ROZDĚLENÍ DOBY NEZAMĚSTNANOSTI V ČESKÉ REPUBLICE

**Ivana Malá***

Nezaměstnanost je velkým problémem všech tržních ekonomik, proto její zkoumání z nejrůznějších úhlů pohledu je velmi důležité. Přes velké úsilí odborníků z nejrůznějších oblastí nebylo nalezeno prakticky fungující řešení, které by přineslo zásadní obrat ve vývoji tohoto nejen ekonomicky, ale také společensky negativního jevu. V současné době je stejně velkým společenským problémem jako míra nezaměstnanosti také doba, po kterou nezaměstnaní novou práci hledají. Dlouhodobě nezaměstnaní tak téměř ztrácejí jakékoliv naděje na nalezení práce. Je dobře známo, že počet nezaměstnaných, počet volných míst a délku nezaměstnanosti ovlivňuje mnoho faktorů. Ekonomové (a jiní odborníci) se snaží hledat tyto faktory, kvantifikovat jejich vliv na nezaměstnanost a případně navrhnout postupy, které by pozitivní vlivy zdůraznily a negativní potlačily.

Uveďme například zásadní vliv ekonomických faktorů, jako jsou celková ekonomická situace, daňová politika, velikost a doba vyplácení dávek v nezaměstnanosti nebo minimální mzda (Daveri, Tabellini, 2000, Krueger et al., 2011, Hunt, 1995, Lechner a kol., 2002, Alba-Ramírez, 1999), demografických faktorů (Löster, Langhamrová, 2011) nebo také zdravotní hledisko (Korpi, 2001). České republiky a Slovenské republiky v devadesátých letech dvacátého století a problémů přechodu ekonomik bývalé „východní“ Evropy se týkají ekonometrické práce Ham a kol., 1998, Burda a kol., 1993 a Svejnar, 2002.

Pro popis rozdělení délky nezaměstnanosti je možné použít neparametrické, semiparametrické a parametrické modely využívající široké spektrum statistických metod a postupů. Zmiňme modely založené na markovských procesech (Lechner a kol., 2002), Coxův regresní model nebo AFT model (Jarošová, 2006, Jarošová, Malá, 2005), logitový model pro modelování pravděpodobností přechodů (Alba-Ramírez, 1999).

V tomto textu je porovnán neparametrický Kaplanův-Meierův odhad funkce přežití (doplňku distribuční funkce rozdělení délky nezaměstnanosti) s parametrickým modelem využívajícím konečnou směs logaritmicko-normálních rozdělení. Jako popis rozdělení doby nezaměstnanosti je zřejmě třeba zvolit doprava zešikmené rozdělení s poměrně těžkým pravým koncem. Co se týče průběhu rizikové funkce (nebo lépe intenzity nacházení práce) je možné volit pravděpodobnostní rozdělení s klesající

* Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky (malai@vse.cz).

rizikovou funkcí nebo s funkcí, která nabývá maxima. Logaritmicko-normální rozdělení má rizikovou funkci rostoucí ke globálnímu maximu a pak pomalu klesající k nule (Jarošová, Malá, 2005). Mimo tohoto rozdělení se používají například logaritmicko-logistické rozdělení, gama rozdělení, Weibullovo rozdělení nebo další rozdělení (McDonald, Butler, 1987, Johnson a kol., 1994).

Pro konstrukci modelů byla využita data z Výběrového šetření pracovních sil (VŠPS), které provádí Český statistický úřad (CZSO). Data kromě velkého množství dalších informací obsahují také údaje o zaměstnanosti a nezaměstnanosti, o délce hledání zaměstnání a další demografické údaje. Pro zařazení nezaměstnaných do skupin byly použity proměnné pohlaví a nejvyšší dosažené vzdělání.

## 1. Metodika

Pro popis doby nezaměstnanosti použijeme analýzu přežití, která se zabývá zkoumáním náhodných veličin, které popisují dobu do určité události. Název analýza přežití pochází z medicíny, kde je často sledována doba do úmrtí nebo doba do znovuobjevení nemoci. Jedná se dále například o dobu bezproblémového splácení úvěru do prvních problémů se splácením, dobu od nahlášení pojistné škody do jejího vypořádání nebo délku soudního řízení. V teorii kontroly a popisu jakosti je sledována například doba do první opravy nebo doba funkčnosti zařízení, v teorii hromadné obsluhy například doba strávená v systému obsluhy. V tomto textu bude touto událostí nalezení nebo znovunalezení práce. Nejprve tedy shrneme základní pojmy, se kterými teorie přežití pracuje. V případě, že předpokládáme rozdělení směsi pravděpodobnostních rozdělení, lze tyto pojmy snadno upravit. Samozřejmě je přirozenou otázkou, kdy lze základní charakteristiky přepsat do tvaru směsi (váženého průměru) stejných charakteristik jejích složek.

Uvažujme náhodnou veličinu $T$ se spojitým rozdělením nabývající pouze nezáporných hodnot. Rozdělení je popsáno hustotou $f(t)$ a distribuční funkcí $F(t)$. Z předpokladu nezápornosti hodnot náhodné veličiny plyne, že $f(t) = F(t) = 0, t \leq 0$. Pro analýzu přežití je výhodnější používat jako charakteristiku pravděpodobnostního rozdělení funkci přežití $S$ (místo distribuční funkce $F$) definovanou v čase $t$ jako pravděpodobnost, že náhodná veličina $T$ nabude hodnoty větší než $t$ (do času $t$ ke sledované události nedojde). Je tedy

$$S(t) = P\left(T > t\right) = 1 - F(t),\ t \in R. \tag{1}$$

Je tedy $S(t) = 1,\ t \leq 0$ a funkce $S$ je spojitá, nerostoucí funkce. Obdobně jako hustota nebo distribuční funkce je tato funkce jednoznačnou charakteristikou rozdělení náhodné veličiny $T$. Kvantily $t_P$ náhodné veličiny $T$ můžeme s pomocí funkce přežití definovat jako řešení rovnice

$$S(t_P) = 1 - P,\ P \in \left(0,1\right), \tag{2}$$

a tedy místo obvyklého $t_P = F^{-1}(P)$ lze použít definici $t_P = S^{-1}(1-P)$. Pokud bychom chtěli určit střední hodnotu $E(T)$ veličiny $T$, platí obdoba výpočtu střední hodnoty z distribuční funkce $F$ ve tvaru

$$E(T) = \int_0^\infty t\ f(t)\,dt = \int_0^\infty S(t)\,dt. \tag{3}$$

Většina rozdělení používaných v analýze přežití není symetrická, proto dáváme často přednost kvantilovým charakteristikám polohy a variability (medián, další vybrané percentily, různá rozpětí) před momentovými charakteristikami, jako jsou střední hodnota a rozptyl. Logaritmicko-normální rozdělení, použité jako model pro doby nezaměstnanosti, je kladně zešikmené, proto je například medián považován za charakteristiku polohy s lepší vypovídací hodnotou, než je střední hodnota ovlivněná řídkými vysokými hodnotami veličiny *T*. V případě doby nezaměstnanosti jde například o nezaměstnané déle než dva roky, kterých bylo (obrázek 2) v roce 2011 téměř 21 %.

Vzhledem k tomu, že budeme zkoumat výskyt události, užitečnou informací je jejich intenzita (intenzita, se kterou události nastávají v určitém čase). Funkci rizika *h*, definujeme jako

$$h(t) = \lim_{\Delta t \to \infty} \frac{P\left(t \le T < t + \Delta t \middle| T \ge t\right)}{\Delta t} = \frac{f(t)}{S(t)},\ t > 0. \tag{4}$$

Lze si ji představit jako pravděpodobnost, že ke zkoumané události dojde bezprostředně po čase *t*, jestliže k ní do tohoto času nedošlo. Čím větší je hodnota rizikové funkce, tím vyšší je intenzita výskytu událostí (tím více nezaměstnaných v tomto období nalezne práci). Pro logaritmicko-normální rozdělení je riziková funkce jednovrcholová mající jedno globální maximum.

Předpokládejme nyní, že sledovaná veličina *T* má rozdělení definované jako konečná směs *K* pravděpodobnostních rozdělení (Titterington a kol., 1985, McLachlan, Peel, 2000) s hustotou ve tvaru

$$f(t;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j f_j\left(t;\boldsymbol{\theta}_j\right), \tag{5}$$

kde váhy $\boldsymbol{\pi}$ splňují podmínky $0 \le \pi_j \le 1$, $\sum_{j=1}^{K} \pi_j = 1$ a dále $f_j(t;\boldsymbol{\theta}_j)$, $j = 1, ..., K$ jsou hustoty pravděpodobnosti jednotlivých složek směsi, které závisejí na *p*-rozměrných vektorech (neznámých) parametrů $\boldsymbol{\theta}_j$. Vektor $\boldsymbol{\psi}$ obsahuje neznámé parametry v modelu, $K-1$ parametrů $\pi_j$, $j = 1, ..., K-1$ a *Kp* složek parametrů komponentních rozdělení $\boldsymbol{\theta}_j$, $j = 1, ..., K$.

Model (5) můžeme použít například v situaci, kdy se zkoumaná populace skládá z *K* podmnožin, v každé má sledovaná náhodná veličina rozdělení s hustotou $f_j(t;\boldsymbol{\theta}_j)$. Často se předpokládá (jako v předkládaném modelu), že všechny hustoty jsou stejné, liší se jen v hodnotách parametrů. Váhy $\boldsymbol{\pi}$ pak představují podíl jednotlivých podmnožin v populaci.

Označme $X_j$, $j = 1, ..., K$ náhodnou veličinu s rozdělením s hustotou $f_j(t;\boldsymbol{\theta}_j)$ a dále $F_j(t;\boldsymbol{\theta}_j)$, $S_j(t;\boldsymbol{\theta}_j)$, $h_j(t;\boldsymbol{\theta}_j)$, $E(X_j)$ a $D(X_j)$ distribuční funkci, funkci přežití, rizikovou funkci, střední hodnotu a rozptyl rozdělení *j*-té komponenty, $j = 1, ..., K$. Nyní tyto charakteristiky zapíšeme pro rozdělení směsi. Z (5) snadno dostáváme

$$F(t;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j F_j\left(t;\boldsymbol{\theta}_j\right), \tag{6}$$

$$S(t;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j S_j\left(t;\boldsymbol{\theta}_j\right),$$

$$E(T) = \sum_{j=1}^{K} \pi_j E\left(X_j\right) = \sum_{j=1}^{K} \pi_j \int_0^{\infty} S_j\left(t;\boldsymbol{\theta}_j\right) dt. \tag{7}$$

Rozptyl směsi je možno určit jako

$$D(T) = E\left(T^2\right) - \left[E\left(T\right)\right]^2 = \sum_{j=1}^{K} \pi_j E\left(X_j^2\right) - \left[\sum_{j=1}^{K} \pi_j E\left(X_j\right)\right]^2.$$

Dále je podle (4) a (5)

$$h(t;\boldsymbol{\psi}) = \frac{\sum_{j=1}^{K} \pi_j f_j\left(t;\boldsymbol{\theta}_j\right)}{\sum_{j=1}^{K} \pi_j S_j\left(t;\boldsymbol{\theta}_j\right)} = \frac{\sum_{j=1}^{K} \pi_j S_j\left(t;\boldsymbol{\theta}_j\right) \dfrac{f_j\left(t;\boldsymbol{\theta}_j\right)}{S_j\left(t;\boldsymbol{\theta}_j\right)}}{\sum_{j=1}^{K} \pi_j S_j\left(t;\boldsymbol{\theta}_j\right)}, \; t > 0 \; \left(\text{pro } S_j\left(t;\boldsymbol{\theta}_j\right) > 0\right). \tag{8}$$

Podle vzorce je riziková funkce opět rovna váženému průměru hodnot rizikových funkcí složek, váhy jsou ale rovny $\pi_j S_j\left(t;\boldsymbol{\theta}_j\right)$, $j = 1, ..., K$ a při tomto zápisu závisejí na hodnotě $t$ a na parametrech $\boldsymbol{\theta}_j$.

Kvantily $t_P$ rozdělení je třeba obecně určit řešením rovnice (2) ve tvaru

$$S\left(t_P\right) = \sum_{j=1}^{K} \pi_j S_j\left(t_P;\boldsymbol{\theta}_j\right) = 1 - P, \; 0 < P < 1. \tag{9}$$

V dalším textu budeme předpokládat, že $f_j$ jsou hustoty dvouparametrického logaritmicko-normálního rozdělení. V takovém případě je $p = 2$, $\boldsymbol{\theta}_j = \left(\mu_j, \sigma_j^2\right)$ a platí

$$\begin{aligned} f(t;\mu,\sigma^2) &= 0, && t \le 0, \\ &= \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln t - \mu)^2}{2\sigma^2}\right) = \frac{1}{t\sigma}\varphi\left(\frac{\ln t - \mu}{\sigma}\right), && t > 0, \end{aligned} \tag{10}$$

a

$$\begin{aligned} S(t;\mu,\sigma^2) &= 0, && t \le 0, \\ &= 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right), && t > 0, \end{aligned} \tag{11}$$

kde $\varphi$ je hustota a $\Phi$ je distribuční funkce normovaného normálního rozdělení. Riziková funkce pro logaritmicko-normální rozdělení roste od hodnoty nula pro $t = 0$ k maximu a dále pomalu klesá opět k nule a lze ji zapsat jako (použijeme (4), (10) a (11))

$$h\left(t;\mu,\sigma^2\right) = \frac{\varphi\left(\frac{\ln t - \mu}{\sigma}\right)}{t\sigma\left[1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)\right]}.$$

Předpoklad logaritmicko-normálního rozdělení pro délku nezaměstnanosti tedy znamená, že předpokládáme, že intenzita získávání práce nejdříve roste, nabývá maxima a pak s časem klesá. Hodnota času, ve kterém riziková funkce nabývá maxima, závisí na odhadnutých parametrech. Riziková funkce směsi logaritmicko-normálních rozdělení nemusí mít tento tvar, v případě odhadnutých směsí v tomto textu tomu tak bude (obrázek 6).

V analýze přežití se běžně setkáváme s neúplnými, tedy cenzorovanými daty. Datové soubory, se kterými je třeba pracovat, obsahují nejen úplná pozorování, kdy máme informaci o času *T*, ve kterém sledovaná událost nastala. Pozorování zprava cenzorovaná znamenají, že známe pouze časový okamžik, do kterého událost nenastala, pro *i*-tou jednotku tedy pouze víme, že $T > T_i$. V případě nezaměstnanosti budou pozorování zprava cenzorovaná pro nezaměstnané, kteří zaměstnání do doby $T_i$ nenašli. Vzhledem k tomu, že v šetření VŠPS respondenti neuvádějí dobu nezaměstnanosti přesně, ale pouze v intervalu (předpokládejme $\left(L_i, U_i\right\rangle\right)$, pro účely odhadu modelu uvažujeme tato pozorování zprava cenzorovaná v čase $L_i$, volíme $T_i = L_i$. V případě cenzorovaných dat pozorujeme dvojice ve tvaru $\left(T_i, C_i\right)$, kde $T_i$ je doba pozorování a $C_i$ je kód cenzorování, pro úplné pozorování budeme volit hodnotu 1, pro zprava cenzorovaná hodnotu 0.

V tomto textu budeme dále uvažovat pozorování intervalově cenzorovaná, neboť šetření VŠPS probíhá po čtvrtletích a budeme vědět, že nezaměstnaný práci našel, nemáme ale informaci o přesné délce nezaměstnanosti. Známe pouze časový interval, ve kterém k nalezení práce došlo. Pro intervalově cenzorovaná data tedy známe interval $\left(L_i, U_i\right\rangle$, ve kterém došlo k výskytu události. V takovém případě nevystačíme s popisem dat pomocí dvojice veličin $\left(T_i, C_i\right)$, použijeme proto trojici, kde pro *i*-té pozorování je $\left(L_i, U_i, C_i\right)$.

Použijeme-li popis pozorování pomocí trojice $\left(L_i, U_i, C_i\right)$, v souladu se značením v programu R budeme uvažovat pro $i = 1, \ldots, n$

- *i*-té pozorování je úplné a k události došlo v čase $T_i$ : $L_i = T_i \left(= U_i\right)$, $C_i = 1$,
- *i*-té pozorování je zprava cenzorované, k události nedošlo do času $T_i$ : $L_i = T_i$, $U_i = \infty$, $C_i = 0$,
- *i*-té pozorování je intervalově cenzorované, k události došlo v intervalu $\left(L_i, U_i\right\rangle$: $L_i, U_i, C_i = 3$.

Hodnota $C$=2 se používá pro zleva cenzorovaná data, která nejsou předmětem tohoto textu.

Pro odhad rozdělení doby nezaměstnanosti je možné použít Kaplanův-Meierův neparametrický odhad funkce přežití. Metoda je založena na principu odhadu pomocí empirické distribuční funkce a tento postup je doplněn o využití cenzorovaných dat (pro zprava cenzorovaná data byla metoda navržena v práci Kaplan, Meier, 1958, úpravu na intervalově cenzorovaná data (používaná v tomto textu) lze najít například v Lawless, 2003). Takový model nevyžaduje žádný předpoklad o pravděpodobnostním rozdělení doby nezaměstnanosti, umožňuje odhad kvantilů rozdělení a testování stejných rozdělení podmnožin. Odhad je konstantní vždy mezi časovými okamžiky $(L_i, U_i, i = 1, ..., n)$ obsaženými v datech seřazenými podle velikosti a v případě intervalů v analyzovaném datovém souboru budou vzdálenosti dlouhé (několik měsíců). Navíc odhad poskytuje informaci o průběhu funkce přežití pouze do poslední (pravé) meze pozorovaných intervalů.

Dále sestrojíme parametrický odhad funkce přežití. Úspěšná aplikace parametrického modelu je podmíněna oprávněností volby modelu rozdělení sledované doby, avšak možnosti posouzení vhodnosti volby modelu jsou v případě dat, která představují pouze cenzorovaná (zprava a intervalově) pozorování, omezené. Určitá nepřesnost při aplikaci modelů na data z VŠPS vzniká v důsledku toho, že nezaměstnaní jsou sledováni (retrospektivně) po nestejnou dobu. V důsledku způsobu výběru z databáze VŠPS mají větší pravděpodobnost zahrnutí do výběru nezaměstnaní s delší dobou trvání nezaměstnanosti (Jarošová, 2006). Potom je třeba počítat s tím, že odhady charakteristik doby trvání získané z těchto dat jsou nadhodnocené.

Neznámé parametry budeme odhadovat metodou maximální věrohodnosti, kvalitu různých modelů porovnáme Akaikovým kritériem. Do věrohodnostní funkce $L$ přispívá $i$-té úplné pozorování hodnotou $f(t_i; \boldsymbol{\psi})$, zprava cenzorované pozorování hodnotou

$$S(t_i) = P(T > t_i) = 1 - F(t_i; \boldsymbol{\psi})$$

a intervalově cenzorované pozorování hodnotou

$$P(l_i < T \le u_i) = F(u_i; \boldsymbol{\psi}) - F(l_i; \boldsymbol{\psi}).$$

Věrohodnostní funkce pak má tvar

$$L(\boldsymbol{\psi}) = \prod_{i:t_i \text{ úplné}} f(t_i; \boldsymbol{\psi}) \prod_{\substack{i:t_i \text{ zprava} \\ \text{cenzorované}}} (1 - F(t_i; \boldsymbol{\psi})) \prod_{\substack{i:t_i \text{ intervalově} \\ \text{cenzorované}}} (F(u_i; \boldsymbol{\psi}) - F(l_i; \boldsymbol{\psi})). \qquad (12)$$

Pro logaritmickou věrohodnostní funkci $l(\boldsymbol{\psi}) = \ln(L(\boldsymbol{\psi}))$ platí

$$l(\boldsymbol{\psi}) = \sum_{i:t_i \text{ úplné}} \ln(f(t_i; \boldsymbol{\psi})) + \sum_{\substack{i:t_i \text{ zprava} \\ \text{cenzorované}}} \ln(1 - F(t_i; \boldsymbol{\psi})) + \sum_{\substack{i:t_i \text{ intervalově} \\ \text{cenzorované}}} \ln(F(u_i; \boldsymbol{\psi}) - F(l_i; \boldsymbol{\psi})). \qquad (13)$$

Pro maximalizaci (12) nebo (13) je třeba použít numerické metody, v případě obecného modelu směsi se používá iterační EM algoritmus (McLachlan, Peel, 2000), který v opakovaných dvou krocích hledá odhad $\hat{\boldsymbol{\psi}}$ neznámého vektoru parametrů $\boldsymbol{\psi}$.

Dále budeme předpokládat, že příslušnost ke složce rozdělení je možné pozorovat. V případě znalosti příslušnosti pozorování ke složce se úloha maximalizace (13) velmi zjednodušuje (Lawless, 2003). Definujme (v tomto případě nenáhodné) *K*-rozměrné vektory $\mathbf{z}_i, i = 1, ..., n$ takové, že

$z_{ij} = 1,$ $i$ – té pozorování pochází z $j$ – té komponenty,

$= 0,$ jinak.

Potom je (podle (5) a (6))

$$f(t_i;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j f_j\left(t;\boldsymbol{\theta}_j\right) = \prod_{j=1}^{K} \left(\pi_j f_j\left(t;\boldsymbol{\theta}_j\right)\right)^{z_{ij}},$$

$$F(t_i;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j F_j\left(t;\boldsymbol{\theta}_j\right) = \prod_{j=1}^{K} \left(\pi_j F_j\left(t;\boldsymbol{\theta}_j\right)\right)^{z_{ij}},$$

a (12) lze přepsat jako

$$L(\boldsymbol{\psi}) = \prod_{i:t_i \text{ úplné}} \prod_{j=1}^{K} \left(\pi_j f_j\left(t;\boldsymbol{\theta}_j\right)\right)^{z_{ij}} \prod_{i:t_i \text{ zprava cenzorované}} \left(1 - \prod_{j=1}^{K} \left(\pi_j F_j\left(t;\boldsymbol{\theta}_j\right)\right)^{z_{ij}}\right) \cdot \cdot \prod_{i:t_i \text{ intervalově cenzorované}} \left(\prod_{j=1}^{K} \left(\pi_j F_j\left(u_i;\boldsymbol{\theta}_j\right)\right)^{z_{ij}} - \prod_{j=1}^{K} \left(\pi_j F_j\left(l_i;\boldsymbol{\theta}_j\right)\right)^{z_{ij}}\right) \tag{14}$$

Pokud zlogaritmujeme (14), lze logaritmickou věrohodnostní funkci $l$ rozdělit na část, ve které odhadneme pravděpodobnosti $\pi_j$, a část, ve které odhadneme v každé složce zvlášť parametry komponentních rozdělení. Lze tedy odhadnout zvlášť pravděpodobnosti $\pi_j$, $j = 1, ..., K$, maximálně věrohodnými odhady jsou $\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij}$ (relativní četnostmi pozorování z *j*-té komponenty v souboru dat) a pro každou komponentu nalézt maximálně věrohodné odhad $\hat{\boldsymbol{\theta}}_j$ parametru $\boldsymbol{\theta}_j$.

Pro modely konstruované v tomto textu vypadne první část věrohodnostní funkce (14), neboť data neobsahují úplná pozorování. Na rozdíl od maximálně věrohodných odhadů parametrů logaritmicko-normálního rozdělení pro úplná data nelze maximálně věrohodné odhady v případě přítomnosti cenzorovaných pozorování v datech zapsat analyticky a je třeba je hledat numericky.

Všechny výpočty byly provedeny v programu R (RPROGRAM). Pro numerické hledání maximálně věrohodných odhadů parametrů rozdělení složek byl použit balíček Survival (RSURVIVAL). Hodnota logaritmické věrohodnostní funkce pro určení hodnoty Akaikova kritéria $\left(AIC = 2 \cdot \text{počet parametrů} - 2 \cdot l\left(\hat{\boldsymbol{\psi}}\right)\right)$ pak byla určena dosazením vektoru $\hat{\boldsymbol{\psi}} = \left(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}_j,\ j = 1, ..., K\right)$ do logaritmické věrohodnostní funkce.

## 2. Data a výsledky

Nyní použijeme výsledky předchozí části pro dobu nezaměstnanosti v České republice, sledovanou událostí tedy bude nalezení (znovunalezení) práce a analyzovanou náhodnou veličinou bude doba nezaměstnanosti (doba hledání zaměstnání). Již bylo uvedeno, že pro analýzu použijeme data z výběrového šetření VŠPS (Výběrové šetření pracovních sil). Šetření provádí čtvrtletně Český statistický úřad od prosince roku 1992. Hlavním cílem VŠPS je získávání pravidelných informací o situaci na trhu práce, umožňujících její analýzu z různých hledisek, zejména ekonomických, sociálních a demografických. Data jsou sbírána prostřednictvím dotazníku a šetření probíhá v domácnostech, výběrovou jednotkou Výběrového šetření pracovních sil je byt. Od roku 2002 jsou obsah a forma dotazníku VŠPS plně harmonizovány se standardem Evropské unie a dotazník je tak národní modifikací celoevropského šetření Labour Force Sample Survey (LFSS).

Šetření, kromě základních informací o bytu a domácnostech v něm žijících, zjišťuje demografické údaje a vazby mezi jednotlivými členy domácností. Nejobsáhlejší částí dotazníku je oddíl zabývající se podrobnými údaji o všech osobách 15letých a starších, obvykle bydlících v bytě (ekonomické postavení, charakteristika hlavního, resp. druhého, zaměstnání, předchozí pracovní zkušenost, hledání zaměstnání, obvyklé postavení, vzdělávání a situace respondenta před rokem). Přístup uplatněný ve VŠPS umožňuje sledovat reálnou situaci domácností a respondentů a vytvářet informační předpoklady pro formulování zásad sociální politiky a politiky zaměstnanosti (CZSO).

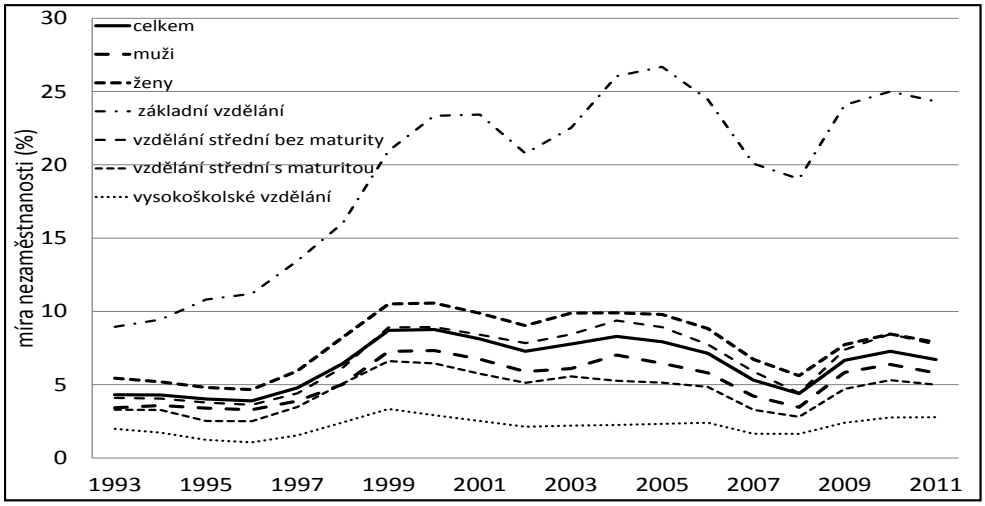
Byty jsou do šetření zařazovány prostřednictvím dvoustupňového výběru. Každý byt zůstává v šetřeném souboru po dobu pěti po sobě jdoucích čtvrtletí, obměna souboru je každé čtvrtletí 20% výběru. Při tomto způsobu rotace jsou získávány konzistentní informace nejen za navazující období, ale šetření umožňuje i porovnání výsledků za respondenta nebo domácnost se stejným obdobím minulého roku. Podle šetření se za nezaměstnané považují všechny osoby patnáctileté a starší, které v průběhu referenčního týdne (týdne konání šetření v daném bytě) nebyly zaměstnané, byly připraveny k nástupu do práce ihned nebo do čtrnácti dnů a v průběhu posledních čtyř týdnů hledaly aktivně práci (CZSO).

Český statistický úřad publikuje čtvrtletně míru nezaměstnanosti, tato čtvrtletní data jsou ovšem ovlivněna sezónními výkyvy. Zmiňme každoroční nárůst nezaměstnanosti v prvním čtvrtletí roku, než na jaře začnou sezónní práce. Na obrázku 1 je znázorněna průměrná roční obecná míra nezaměstnanosti pro všechny nezaměstnané, a dále tato hodnota zvlášť pro muže a ženy (silné čáry) a pro skupiny nezaměstnaných (bez ohledu na pohlaví) popsané nejvyšším dosaženým vzděláním. Na obrázku je zřetelně patrná vysoká nezaměstnanost osob, které mají pouze základní vzdělání

a nízká nezaměstnanost osob se vzděláním vysokoškolským. Míra nezaměstnanosti žen (bez ohledu na vzdělání) je v celém sledovaném období vyšší než míra nezaměstnanosti osob se středním vzděláním bez maturity. Všimněme si dále, že nezaměstnanost osob se středním vzděláním s maturitou velmi dobře kopíruje celkovou nezaměstnanost v České republice.

**Obrázek 1**

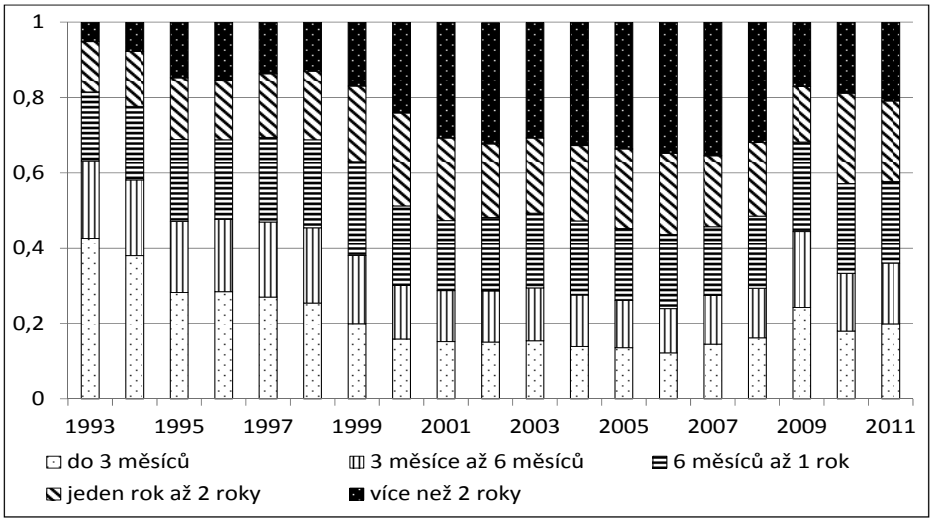
**Obecná míra nezaměstnanosti v letech 1993–2011**




Pramen: Český statistický úřad.

Z obrázku je patrné, že vývoj je obdobný pro všechny zkoumané skupiny a liší se posunutím (a v případě základního vzdělání /horní křivka/ také velikostí kolísání). Nejnižší křivka znázorňující nezaměstnané s terciárním vzděláním je v podstatě konstantní. Předmětem zkoumání v tomto článku je ovšem délka nezaměstnanosti, nikoliv pouze její procento. Statistický úřad publikuje počty nezaměstnaných v intervalech 0–3 měsíce, 3–6 měsíců, 6 měsíců až jeden rok, jeden až dva roky a více než dva roky (CZSO). Pokud nezaměstnaný hledá práci déle než jeden rok (dvě poslední třídy zmíněného dělení), patří mezi dlouhodobě nezaměstnané. Dlouhodobou nezaměstnaností, velkým problémem rozvinutých ekonomik, se z pohledu krajů České republiky z demografického hlediska zabývá práce Löster, Langhamrová, 2011. Na obrázku 2 je znázorněn vývoj procentního zastoupení uchazečů o zaměstnání v jednotlivých skupinách, opět od roku 1993. V letech 2010 a 2011 bylo dlouhodobě nezaměstnaných 42 procent, zatímco v letech 2000–2008 bylo dlouhodobě nezaměstnaných přes 50 procent nezaměstnaných. Tento text se dále zabývá naopak dobou nezaměstnanosti pro osoby, které jsou nezaměstnané do dvou let.

**Obrázek 2**
**Rozložení uchazečů o zaměstnání podle délky hledání práce v letech 1993–2011**




Pramen: Český statistický úřad.

Pro analýzu byla použita data o všech nezaměstnaných, kteří byli zahrnuti do pěti po sobě následujících šetření VŠPS prováděných od prvního čtvrtletí roku 2010 do prvního čtvrtletí 2011. V předchozím textu bylo uvedeno, že Český statistický úřad publikuje počty nezaměstnaných v intervalech 0–3 měsíce, 3–6 měsíců, 6 měsíců až jeden rok, jeden až dva roky a více než dva roky. Ve výběrovém šetření VŠPS jsou nezaměstnaným nabízeny pro jejich délku nezaměstnanosti intervaly do jednoho měsíce, 1–3 měsíce, 3–6 měsíců, 6–12 měsíců, 1–2 roky, 2– 4 roky a déle než čtyři roky.

Údaje šetření VŠPS neobsahují přesné délky nezaměstnanosti (například ve dnech nebo týdnech), lze nalézt informaci o tom, zda nezaměstnaný během pěti čtvrtletí, po která byt nezaměstnaného zůstává v šetřených bytech, práci nalezl nebo nenalezl (nebo také ztratil a následně nalezl nebo nenalezl). Z dostupných údajů lze sestrojit interval pro dobu nezaměstnanosti v případě, že nezaměstnaný práci nalezl (intervalově cenzorovaná pozorování) a dále dobu, po kterou je již nezaměstnaný bez práce v případě, že práci nenalezl (zprava cenzorovaná data). Po úpravě pomocí dalších údajů (například zpoždění nástupu práce) byly určeny dolní meze intervalů cenzorování $l$ (v měsících) 0, 1, 3, 4, 6, 9, 12 a 18 a horní meze $u$ 1, 3, 4, 6, 9, 12, 15, 18, 21, 24 a 27 měsíců.

V další části sestrojíme model rozdělení doby nezaměstnanosti jako směs logaritmicko-normálních rozdělení. Vzhledem k tomu, že při použití takového modelu je důležitá volba pravděpodobnostního rozdělení, je sestrojen také neparametrický model, který žádný takový předpoklad nevyužívá. Například v aplikacích v lékařství se Kaplanovu-Meierovu modelu dává přednost před parametrickým modelem, použití parametrického modelu v případě vhodně zvoleného modelového rozdělení přináší

výhody širokého spektra metod parametrické statistiky. Na druhé straně pro nevhodné rozdělení můžeme získat zavádějící nebo naprosto špatné výsledky.

Budeme uvažovat komponenty dané pohlavím nezaměstnaného (směs dvou rozdělení pro muže a ženy) a dále nejvyšším dosaženým vzděláním (směs tří rozdělení pro komponenty základní vzdělání a středoškolské bez maturity, středoškolské s maturitou a vysokoškolské vzdělání). Třídy základní vzdělání nebo bez vzdělání byly spojeny se středoškolským vzděláním bez maturity, neboť v datech nebylo možné nalézt tolik osob bez vzdělání, které našly práci, aby bylo možné odhadovat parametry rozdělení. Zvolený model obsahuje v prvním případě (obecně) pět parametrů (1+4) a ve druhém osm parametrů (2+6). Do modelu byli zařazeni všichni nezaměstnaní ve věku 16–65 let, kteří práci našli do 24 měsíců, nebo jsou nezaměstnaní do 24 měsíců. Průměrný věk 4 753 nezaměstnaných byl 37,5 roku. Pokud nezaměstnaný po dobu sledování nalezl zaměstnání a zase ho ztratil, byl započítán pouze jednou jako nezaměstnaný, který nalezl zaměstnání. Žádný nezaměstnaný, který by nalezl (a ztratil) ve sledované době zaměstnání dvakrát, nalezen nebyl.

Data obsahují také informaci o tom, zda je nezaměstnaný registrován na úřadu práce a pokud ano, zda pobírá nebo nepobírá podporu v nezaměstnanosti. V analyzovaném souboru nezaměstnaných je 61 % registrovaných uchazečů o zaměstnání a z nich pouze jedna třetina pobírá dávky v nezaměstnanosti. Z předchozího je zřejmé, že data pocházející z šetření VŠPS jsou jiná než data pocházející z registrů Úřadů práce a Ministerstva práce a sociálních věcí (MPSV). Z dat získaných v rámci VŠPS se určuje obecná míra nezaměstnanosti, z údajů MPSV pak registrovaná míra nezaměstnanosti. Obě míry nezaměstnanosti pak pravidelně publikuje Český statistický úřad (CZSO).

Datový soubor neobsahuje úplná pozorování, logaritmická věrohodnostní funkce má proto tvar

$$l(\boldsymbol{\psi}) = \sum_{i:t_i \text{ zprava cenzorované}} \ln\left(1-\prod_{j=1}^{K}\left(\pi_j F_j\left(t;\boldsymbol{\theta}_j\right)\right)^{z_{ij}}\right) \cdot \sum_{i:\ \substack{\text{intervalově}\\ \text{cenzorované}}} \ln\left(\prod_{j=1}^{K}\left(\pi_j F_j\left(u_i;\boldsymbol{\theta}_j\right)\right)^{z_{ij}} - \prod_{j=1}^{K}\left(\pi_j F_j\left(l_i;\boldsymbol{\theta}_j\right)\right)^{z_{ij}}\right).$$

Výše popsané modely označíme jako

I. Jedno logaritmicko-normální rozdělení, dva parametry $\left(\mu_1, \sigma_1^2\right)$,
II. Dvě komponenty definované pohlavím nezaměstnaného, logaritmicko-normální rozdělení komponent, 5 parametrů. Na základě analýzy dat byl zvolen model, který předpokládá stejné parametry rozptylu logaritmu doby nezaměstnanosti, a tedy ve skutečnosti odhadujeme čtyři parametry $\left(\pi_1, \mu_1, \mu_2, \sigma^2\right)$,
III. Tři komponenty definované nejvyšším dosaženým vzděláním, logaritmicko-normální rozdělení komponent, osm parametrů $\left(\pi_1, \pi_2, \mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2\right)$.

Vzhledem k tomu, že data obsahují také postavení v předcházejících šetřeních, lze v nich nalézt také údaje nezaměstnaných, kteří v prvním čtvrtletí roku 2010 absolvovali například již pátou návštěvu. Částečně tedy data obsahují omezenou informaci

až o jeden rok dozadu. Vzhledem ke krátkému časovému období nebyla do modelu zařazena sezónní složka, i když je známo, že se čtvrtletí liší jak v míře nezaměstnanosti, tak v šanci práci najít.

V tabulce 1 jsou uvedeny odhadnuté parametry rozdělení pravděpodobností jednotlivých komponent a dále odhady střední hodnoty a mediánu těchto komponentních rozdělení. Všechny sledované modely poskytují velmi podobné rozdělení směsi a tím také charakteristiky z tohoto rozdělení odvozené. Vážený průměr středních hodnot z tabulky 1 (podle (7)) poskytuje odhadnutou střední dobu nezaměstnanosti 22 měsíců pro model I, 21,8 měsíce pro model II a 21,9 měsíce pro model III. Mediány je třeba najít numericky řešením rovnice (9), pro všechny modely dostáváme 14 až 14,1 měsíce, tedy dobu delší než jeden rok. Všimněme si, že jediná komponenta tvořená nezaměstnanými vysokoškoláky má medián doby nezaměstnanosti menší než jeden rok (10,7 měsíce). Z tabulky 1 je patrný velký rozdíl mezi mediány a středními hodnotami. V tomto případě je možno považovat medián za charakteristiku s větší vypovídací hodnotou. V grafu 3 je sestrojen také neparametrický odhad funkce přežití. Medián doby hledání práce (nebo i jiné kvantily) je možné odhadnout také z tohoto odhadu. Pro celý soubor byl nalezen odhad mediánu 13,5 měsíce, což je o půl roku kratší doba než v parametrickém modelu. Pro zkoumané podmnožiny jsou odhadnuté mediány shodné (a rovné 13,5 měsíce) pro muže a vzdělání středoškolské a vyšší. Hodnota 19,5 měsíce byla nalezena pro skupinu nezaměstnaných žen a pro nezaměstnané se základním vzděláním.

**Tabulka 1**

**Odhady parametrů a charakteristik polohy (střední hodnota, medián v měsících) pro komponenty směsí (modely I–III)**

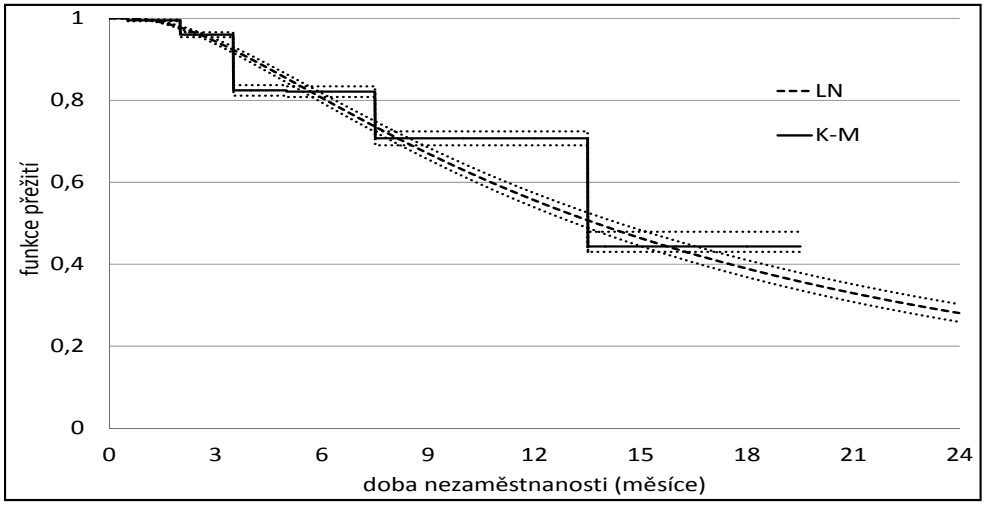
| model | komponenta | *n* | $\mu$ | $\sigma$ | $\pi$ | střední hodnota | medián |
|---|---|---|---|---|---|---|---|
| II | muži | 2 352 | 2,588 (0,029) | 0,937 (0,020) | 0,495 | 20,6 | 13,3 |
| | ženy | 2 401 | 2,703 (0,030) | 0,937 (0,020) | 0,505 | 23,2 | 14,9 |
| III | Z+SŠ | 2 959 | 2,736 (0,030) | 0,937 (0,026) | 0,623 | 23,9 | 15,4 |
| | SŠ + mat. | 1 447 | 2,511 (0,038) | 0,907 (0,034) | 0,304 | 18,6 | 12,3 |
| | VŠ | 347 | 2,371 (0,079) | 0,958 (0,034) | 0,073 | 16,9 | 10,7 |
| I | celkem | 4 753 | 2.644 (0,023) | 0,946 (0,020) | 1 | 22,0 | 14,1 |

Pramen: Vlastní výpočty, Český statistický úřad.

Variabilita délky nezaměstnanosti (měřená směrodatnou odchylkou nebo kvartilovou odchylkou) je nejmenší pro skupinu nezaměstnaných s vysokoškolským vzděláním a pro nezaměstnané muže. Větší proměnlivost je pro skupinu nezaměstnaných žen a skupinu středoškoláků s maturitou, největší variabilita je u skupiny nezaměstnaných, kteří mají maximálně střední vzdělání bez maturity. Celková variabilita je srovnatelná s hodnotou pro nezaměstnané ženy a pro skupinu nezaměstnaných středoškoláků. V případě použití takovéto směsi neplatí, že by komponentní rozdělení měla menší variabilitu než rozdělení všech nezaměstnaných.

Kvalitu odhadů můžeme porovnat pomocí Akaikova kritéria, které umožňuje také zohlednit různý počet odhadovaných parametrů. Komponenty jsou v prezentovaném modelu voleny na základě zvolených vysvětlujících proměnných a ne tak, aby co nejlépe (ve smyslu co největší hodnoty věrohodnostní funkce) popisovaly data, jak tomu je při konstrukci umělých složek (Lawless, 2003). Přesto použití komponent umožňuje snížení hodnoty Akaikova kritéria, pokud jsou skupiny vhodně zvoleny. Hodnoty Akaikova kritéria jsou 6 038 pro model I, 6 032 pro model II a 6 008 pro model III. Nejmenší hodnoty tedy nabývá model směsi s komponentami danými vzděláním nezaměstnaného. Výrazný pozitivní vliv vzdělání na délku nezaměstnanosti je známý, je patrný také na obrázku 5. Je také známo, že doba nezaměstnanosti závisí na pohlaví žadatele o práci, model konstruující směs dvou komponentních rozdělení podle pohlaví umožňuje konstruovat dvousložkový model s odlišnými středními hodnotami a stejnými rozptyly logaritmů doby nezaměstnanosti. Vzhledem k tomu, že výpočet střední hodnoty (na rozdíl od mediánu) i rozptylu logaritmicko-normálního rozdělení vyžaduje znalost obou parametrů, sledovaný model uvažuje různé rozptyly i střední hodnoty doby nezaměstnanosti. Na obrázcích 3–5, porovnáním neparametrických a parametrických křivek, získáváme velmi podobné výsledné křivky pro oba přístupy.

**Obrázek 3**
**Kaplanův-Meierův odhad a parametrický odhad funkce přežití (model I)**

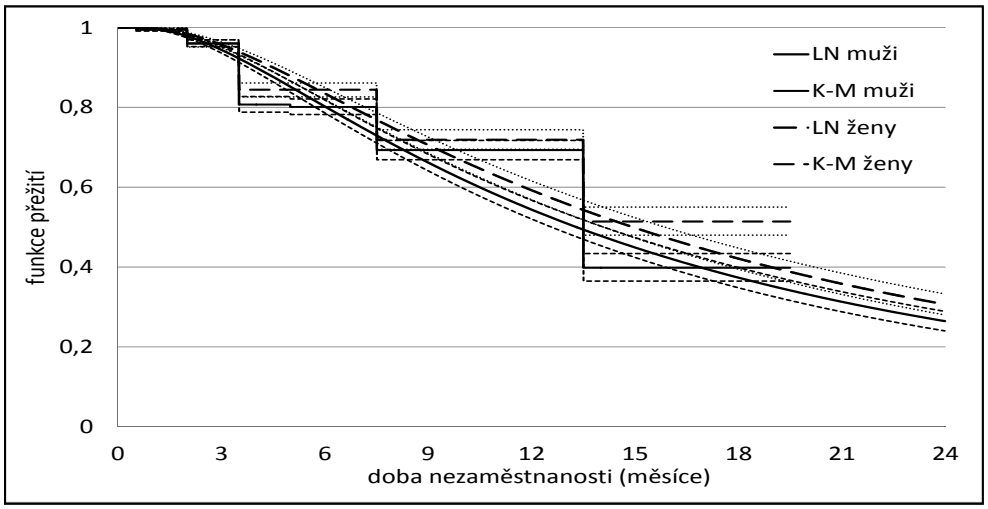



Pramen: Vlastní výpočty, Český statistický úřad.

Na obrázku 3 je znázorněn Kaplanův-Meierův neparametrický odhad funkce přežití spolu s odhadem získaným proložením logaritmicko-normálního rozdělení (maximálně věrohodný odhad). Oba odhady jsou doplněny intervaly spolehlivosti. Všimněme si, že zatímco Kaplanův-Meierův odhad poskytuje odhad funkce *S* jako po částech lineární funkci, a to pouze do 20 měsíců, parametrický odhad je konstruován dosazením do známého teoretického vztahu a možné prodloužit i pro hodnoty nad

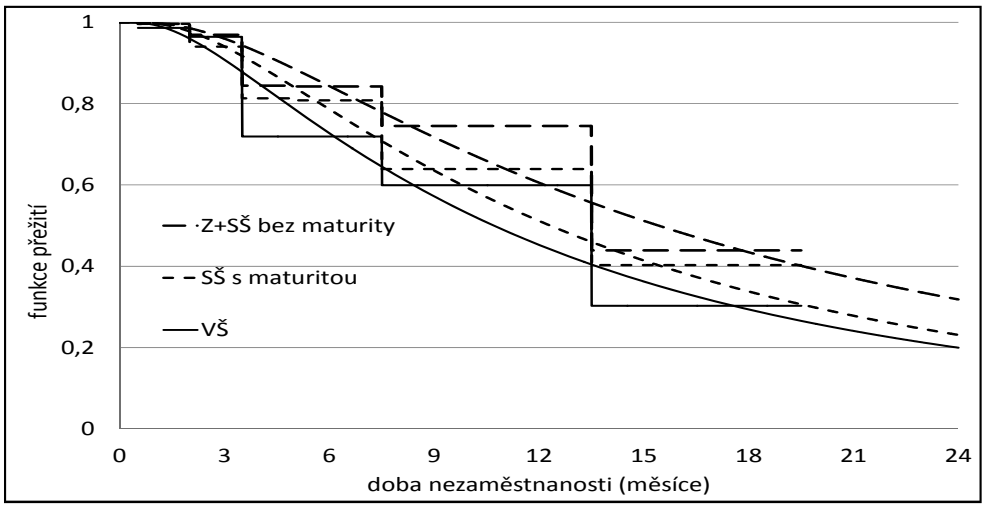
sledovaných 24 měsíců. Z maximálně věrohodných odhadů parametrů lze vyčíslit maximálně věrohodné odhady jakýchkoliv potřebných charakteristik sledovaného rozdělení. V případě modelu směsi máme takovou informaci o zvolených komponentách a o jejich vztahu k charakteristikám celého základního souboru.

**Obrázek 4**
**Kaplanův-Meierův odhad a parametrický odhad funkce přežití (model II)**




Pramen: Vlastní výpočty, Český statistický úřad.

**Obrázek 5**
**Kaplanův-Meierův odhad a parametrický odhad funkce přežití (model III)**

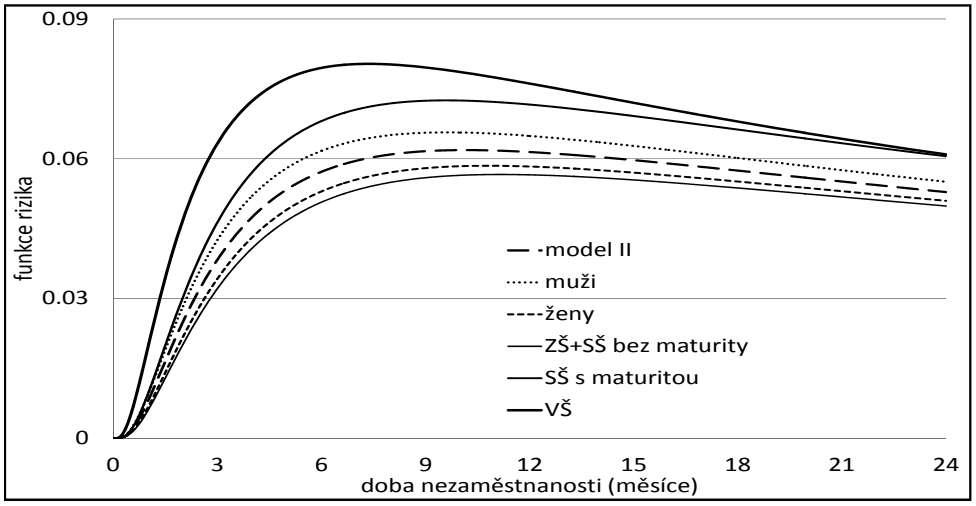



Pramen: Vlastní výpočty, Český statistický úřad.

Obrázky 4 a 5 znázorňují parametrické a neparametrické odhady funkcí přežití pro muže a ženy (model II, obrázek 4) a pro skupiny popsané nejvyšším dosaženým vzděláním (model III, obrázek 5). V případě obrázku 5, který obsahuje šest odhadnutých křivek, již nejsou uvedeny intervaly spolehlivosti. Na obrázku 4 je patrné rychlejší nacházení práce pro muže, na konci dvouletého období je rozdíl $0{,}514 - 0{,}398 = 0{,}116$, tedy mužů našlo zaměstnání o 11,6 procentního bodu více než žen. Na obrázku 5 pak je zřejmý pozitivní vliv vzdělání na délku nezaměstnanosti, rozdíl mezi neparametrickými křivkami pro vysokoškolské vzdělání a pro základní a středoškolské vzdělání je 10 procentních bodů. Obdobné rozdíly vidíme také u parametrického modelu.

Ukažme ještě (na obrázku 6) průběh rizikových funkcí pro zkoumané komponenty českých domácností. Z modelů směsí je zařazen model II (čerchovaná čára), model III by poskytl rizikovou funkci na grafu nerozlišitelnou. Maximální hodnoty těchto funkcí jsou od 7,4 měsíce pro nezaměstnané s vysokoškolským vzděláním do 11,1 měsíce pro nezaměstnané se vzděláním do středoškolského bez maturity. Pro model směsi II dosahuje funkce rizika maxima pro 10,2 měsíce.

**Obrázek 6**
**Odhad rizikové funkce pro uvažované modely**




Pramen: Vlastní výpočty, Český statistický úřad.

## Závěr

V předchozím textu byl popsán model, který na základě dat z Výběrového šetření pracovních sil pořádaného Českým statistickým úřadem umožňuje popsat rozdělení doby nezaměstnanosti v České republice v roce 2010. Byl porovnán neparametrický a parametrický model. Parametrický model umožňuje na základě odhadnutých parametrů a jejich kovarianční matice konstruovat odhady nejrůznějších charakteristik, pro které známe výrazy určené na základě známých vlastností rozdělení. Správná aplikace parametrického modelu ovšem předpokládá vhodně zvolené pravděpodob-

nostní rozdělení. Při volbě byly využity požadované vlastnosti rozdělení, rozdělení použitá v literatuře a implementovaná ve statistických programech a také srovnání s neparametrickým odhadem, který na předpokladu rozdělení nezávisí.

Metoda konečných směsí s pozorovatelnými příslušnostmi ke složkám umožňuje kromě popisu rozdělení doby nezaměstnanosti pro všechny nezaměstnané získat také informace o jednotlivých komponentách (době nezaměstnanosti žen, mužů, nezaměstnaných s daným nejvyšším dosaženým vzděláním).

Výsledky předkládané analýzy ukazují známé a běžně uváděné závislosti míry a délky nezaměstnanosti na pohlaví a vzdělání respondenta, dovolují však také rozdíly kvantifikovat (a případně testovat). Bylo zvoleno porovnání pomocí grafického znázornění funkcí přežití a rizika a výpočtu charakteristik polohy a variability, na základě parametrického modelu by bylo možné vyčíslit i jiné zajímavé veličiny.

Odhady na základě použitých dat z Výběrového šetření pracovních sil lze snadno numerickými metodami získat, problémem je ovšem, že data jsou silně cenzorovaná (100 procent cenzorovaných, 70 % zprava cenzorovaných pozorování nezaměstnaných, kteří práci ve sledovaném období nenašli) a intervaly cenzorování jsou i po všech úpravách a po využití dostupné informace z šetření dlouhé.

## Literatura


ALBA-RAMÍREZ, A. Explaining the Transitions out of Unemployment in Spain: the effect of unemployment insurance. *Applied Economics.* 1999, vol. 31, s. 183–193.

BURDA, M.; BEAN C.; SVEJNAR, J. Unemployment, Labour Markets and Structural Change in Eastern Europe. *Economic Policy.* 1993, vol. 8, no. 16, s. 101–137.

ČABLA, A. Unemployment duration in the Czech Republic. In *International Days of Statistics and Economics at VŠE, Prague, 13. 09. 2012 – 15. 09. 2012.* Praha : VŠE, 2012, s. 257–267.

DAVERI, F.; TABELLINI, G. Unemployement and taxes – do taxis affect the rate of unemployement? *Economic Policy.* 2000, vol. 30, s. 47–88.

HAM, J. C.; SVEJNAR, J.; TERRELL, K. Unemployment and the Social Safety Net during Transitions to a Market Economy: Evidence from the Czech and Slovak Republics. *The American Economic Review.* 1998, vol. 88, no. 5, s. 1117–1142.

HUNT, J. The Effect of Unemployment Compensation on Unemployment Duration in Germany. *Journal of Labor Economics*. 1995, vol. 13, no. 1, s. 88–120.

JAROŠOVÁ, E. Modelování délky trvání nezaměstnanosti. *Statistika.* 2006, roč. 86, č. 3, s. 240–251.

JAROŠOVÁ, E.; MALÁ, I. Modelling time of unemployment in the Czech Republic. APLIMAT 2005 – 4th international conference, Proceedings, s. 465–470.

JOHNSON, N. L.; BALAKRISHNAN, N.; KOTZ, S. *Continuous Univariate Distributions*. Vol. 1., Vol 2. New York : John Wiley, Sons, 1994.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assn.* 1958, vol. 53, s. 457–481.

KORPI, T. Accumulating Disadvantage: Longitudinal Analyses of Unemployent and Physical Health in Representative Samples of the Swedish Population. *European Sociological Review.* 2001, vol. 17, no. 3, s. 255–273.

KRUEGER, A. B.; MUELLER, A.; DAVIS, S. J.; AY´EGUL ´AHIN. Job Search, Emotional Well-Being, and Job Finding in a Period of Mass Unemployment: Evidence from High Frequency Longitudinal Data [with Comments and Discussion]. *Brookings Papers on Economic Activity*, 2011. s. 1–81.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. 2. ed. Hoboken : John Wiley, Sons, 2003.

LECHNER, M.; PUHANI, P. A.; DJURDJEVIC, D. Microeconometric. Analyses of the Structure and Dynamics of Swiss Unemployment. Second Interim Report on the NFP 4045 – 59673 Project. 2002.

LÖSTER, T.; LANGHAMROVÁ, J. Analysis of Long-term Unemployment in the Czech Republic. In LÖSTER, T.; PAVELKA (ed.). *International Days of Statistics and Economics, Praha 22. – 23. 12. 2011.* Slaný : Melandrium, 2011, s. 228–234.

MCDONALD, J. B.; BUTLER, R. J. Some Generalized Mixture Distributions with an Application to Unemployment Duration. *The Review of Economics and Statistics*. 1987, vol. 69, no. 2, s. 232–240.

MCLACHLAN, G. J.; PEEL, D. *Finite Mixture Models.* Wiley series in Probability and Mathematical Statistics: Applied Probability and Statistics Section. New York, 2000.

SVEJNAR, J. Transition Economies: Performance and Challenges. *The Journal of Economic Perspectives*. 2002, vol. 16, no. 1, s. 3–28.

TITTERINGTON, D. M.; SMITH, A.F.; MAKOV, U. E. *Statistical analysis of finite mixture distributions*. Wiley, Sons, 1985.

**Internetové zdroje:**

CZSO. Český statistický úřad. www.czso.cz. 1. 4. 2013.

RPROGRAM. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. www.R-project.org.

RSURVIVAL. Therneau T. *A Package for Survival Analysis in S*. R package version 2.37-4, 2013. http://CRAN.R-project.org/package=survival.

## THE USE OF FINITE MIXTURES OF PROBABILITY DISTRIBUTIONS FOR MODELLING THE DISTRIBUTION OF THE DURATION OF UNEMPLOYMENT IN THE CZECH REPUBLIC

**Abstract:** Unemployment belongs to the most serious economic and social problems of developed countries. Usually, the rate of unemployment is analysed. Another problem is the duration of unemployment and especially long-term unemployment. The unemployment duration in the Czech Republic in 2010 is analysed in the paper. The model uses data from the Labour Force Sample Survey, which is performed quarterly by the Czech Statistical Office. The probability distribution of unemployment duration is modelled with the use of finite mixtures of lognormal distributions with the observable components of membership, gender and education. The observations are right and interval-censored, exact values of the unemployment duration are not included in the data. Both parametric and non-parametric Kaplan-Meier methods are used to estimate the survival function. The estimated survival functions are compared graphically and medians are evaluated for each component. A positive effect of education on the duration of unemployment is found. Also, a greater median unemployment duration is found for women than for men. All the computations are made in the R software.

**Keywords:** unemployment duration, censored data, mixture of probability distributions, survival analysis, Kaplan-Meier estimator

**JEL Classification:** C41, J64