Environmental and Ecological Statistics (2022) 29:587–606
https://doi.org/10.1007/s10651-022-00536-5

New methods of life expectancy estimation

Phuc Ho Dang[1] · Thanh Nga Nguyen[1,2]

Received: 25 September 2021 / Revised: 16 February 2022 / Accepted: 24 March 2022 /
Published online: 13 May 2022

Abstract
Two novel methods of life expectancy estimation, applied to various annual reported demographic datasets, are proposed. First, for datasets that fully recorded birth date and death date of all dead individuals, we rely on the well-known Kaplan–Meier estimation method to provide an accurate estimation framework of life expectancy. Our proposed method can be used as a gold standard in the accuracy investigation of other life expectancy estimation methods. The method can be applied for small areas, where complete mortality data are regularly produced by routine annual surveys. The second new created method, called as local parametric method, based on the theoretical background of survival process with local parametric Weibull distributions, estimates life expectancy using abridged survival data. Experiments on real longitudinal datasets show the new method provides very exact life expectancy estimations for 10 among 15 one-year datasets, whilst the method of Chiang often yields overestimations.

Keywords Chiang method of estimation · Kaplan–Meier estimation · Small areas · Survival analysis · Weibull distribution

1 Introduction

Life expectancy, usually understood as human average life time, has been used as a measure of the health status of the population of England and Wales since the 1840's. For example, in 1841 life expectancy for men in Surrey was 44 years, compared to 25

Handling Editor: Luiz Duczmal.

✉ Phuc Ho Dang
hdphuc@math.ac.vn

Thanh Nga Nguyen
ngant@hvnh.edu.vn

[1] Institute of Mathematics, VAST, 18 Hoang Quoc Viet, Cau Giay, Hanoi 11307, Vietnam

[2] Banking Academy, Chua Boc, Hanoi 11518, Vietnam

years for men in Liverpool (see HMSO 1843). At the same time, length of life was used by William Farr to assess the health of populations and to make international comparisons between countries health (see Eyler 1979). Life expectancy, nowadays obtained using modern mathematical tools of statistics, is selected by researchers as an indicator to examine geographic and socio-demographic inequalities in mortality (see Griffiths and Fitzpatrick 2001, for instance).

Estimation of life expectancy requires to construct a life table to record the proportion alive at each age. The true average lifespan would need to follow up for over 100 years and such cohort specific life tables would be impractical for routine purposes. However, with a current demographic life table, we can define life expectancy at birth as the average life length of a newborn child, if current age specific mortality rates are or will be applied in the future. Life expectancy at birth in an area can therefore be determined as an estimate of the average number of years a new-born baby would survive, if he or she experienced the particular area connected to age-specific mortality rates for that time period throughout his or her life. There have been a rich literature in the area of current life tables (e.g. King 1914; Reed and Merrell 1932; Greville 1943; Keyfitz 1966, 1970). However, statisticians have generally found these methods less than satisfactory because of the lack of clear and direct theoretical motivation. Then, (Chiang 1972, 1978, 1984) proposed a new method of current life tables creation based on survival theory, which is quite simple and can be widely applicable (see Manton et al. 1991; Griffiths and Fitzpatrick 2001; Cristia 2009; Kaplan et al. 2014; OECD 2017).

The Chiang method provides abridged life tables that aggregate deaths and population data into age groups under 1, 1–4, 5–9 ... 80–84, 85 and over. The aggregated deaths and population data in the age groups yield age-specific mortality rates used to calculate the probability of dying at each age interval. These probabilities are then applied to a hypothetical population cohort of newborn babies, to create the formula of life expectancy estimation. Nevertheless, it has been shown in the works Pollard (1989), Hsieh (1991) and Wilmoth et al. (2007), that there are biases in Chiang's current life tables related to the estimations of age specific mortality rates M_x and of the conditional probabilities of dying q_x in age intervals $[x; x+n)$. For instance, the formula $M_x = D_x/P_x$, where D_x and P_x are the number of deaths and the population in the age interval, might provide an incorrect estimation of age specific death rate. This is because parts of D_x and of P_x might belong to the $(x+1)$ age group, not just to the x age group. The above potential misplacement might impact the accuracy of the Chiang life expectancy estimation. On the other hand, the ratio of the deaths' amount and population number in a 5-years age interval usually provides a bias estimation of age-specific mortality rates M_x (see Golbeck 1986; Pollard 1989; Meulen 2012), even when the above misplacement is absent. Consequently, the probability q_x of dying at the age interval calculated from M_x is also biased. Besides, as it is discussed in many studies (see Silcocks et al. 2001; Toson and Baker 2003; Eayres and Williams 2004; Silcocks 2004), the Chiang method of life expectancy estimation has quite low performance in small areas' considerations.

Given the above, this article aims to develop novel estimation methods to address the mentioned lacks in the Chiang method. We then evaluate the accuracy of the proposed method and the method of Chiang by comparing their estimation results with those of the Kaplan–Meier estimation method. The rest of the paper is organized as follows. Section 2 presents a method of life expectancy estimation based on the ordinary Kaplan–Meier estimation method, that can be applied to fully observed longitudinal data. The maximally extracted information with exact ages of all deaths from datasets allows one to create a very exact estimation of life expectancy. The result of that estimation method can be treated as a gold standard to investigate the accuracy of other estimations that rely on abridged data containing only pairs of deaths number and population in each 5 years age interval. The Kaplan–Meier estimation method is a good tool for mortality dataset of small areas, where death events dates were completely recorded. Then the comparison between life expectancies of small areas can be done by using the log—rank test for checking the difference between two Kaplan–Meier survival functions.

In Sect. 3, we propose a novel method of life expectancy estimation based on theoretical background of survival process with local parametric Weibull distributions. Our new method, called ”Local parametric method” (LPM), can be applied to abridged datasets instead of the Chiang method. In Sect. 4, we experiment the life expectancy estimation on a longitudinal survival dataset of FilaBavi (see Nguyen and Vinod 2003) using the Kaplan–Meier estimation method, the Chiang method, and our proposed LPM method. We then compare the obtained results to study the accuracy of these methods. In Sect. 5, a brief conclusion is presented summarizing main results obtained in the previous sections. Then we suggest to apply our proposed LPM method to estimate life expectancy having an abridged demographic dataset.

2 Kaplan–Meier method of life expectancy estimation

In 1958, Kaplan and Meier have proposed a crucially important tool to estimate survival function when only an incompletely observed dataset is given. The well known Kaplan–Meier method is very effective to estimate survival function and hence is widely used in various theoretical as well as application studies.

Let T be a nonnegative random variable that represents the death time of an individual from a homogeneous population $\mathcal{P}$. The survivor function (also known as the survival function) of T is defined as

$$S(t) := \mathbb{P}\{T > t\}\,, \;\; t \in [0; \infty)\,.$$

In other words, $S(t) = 1 - F(t)$, where F denotes the cumulative distribution function of random variable T. The life expectancy (LE) of an individual from the population $\mathcal{P}$ can be found equal to the expectation of the variable T, i.e. $LE = \mathbb{E}[T]$. It is clear that

$$LE = \int\limits_0^\infty t\, dF(t) = \int\limits_0^\infty t\, \frac{-dS(t)}{dt}\, dt\,. \tag{2.1}$$

Thus, life expectancy can be estimated indirectly by using an estimated survival function. In the sequel, we discuss on how to extend the Kaplan–Meier method to obtain a good estimation of life expectancy according to the idea of formula (2.1). Inspired by the original method of survival function estimation, we continue to use the term ”Kaplan–Meier estimation” for the case of life expectancy estimation.

2.1 Kaplan–Meier life expectancy estimation for cohort data

In a fully observed survival cohort dataset, each individual from a population of the size N is fully observed from its birthday to the death moment at the age of t_x, $x = 1, 2, ..., N$. Then the life expectancy of the population equals to the average life span of all individuals in this population:

$$LE = \frac{t_1 + t_2 + ... + t_N}{N}\,. \tag{2.2}$$

In a partially observed survival cohort dataset, where there are censors, we can denote the ordered age times by $t_{(1)} < t_{(2)} < ... < t_{(I)}$ when events (deaths or censors) occur. Here I is the total number of separate age moments. Let d_x and c_x denote the number of deaths and the number of censors occurred at the age moment $t_{(x)}$. Let n_x be the number of individuals remained in population just before the age moment $t_{(x)}$. We then have $n_0 = N$, $n_x = n_{x-1} - d_{x-1} - c_{x-1}$ for $x = 1, 2, ..., I$. The Kaplan–Meier estimation (see Kaplan and Meier 1958) of survival function $S(t)$ is equal

$$\hat{S}(t) = \prod_{t_{(x)} \leq t} \frac{n_x - d_x}{n_x}\,.$$

It is clear that $\hat{S}(t)$ is a positive step-down function and the quantity $p_x = \hat{S}(t_{(x)}) - \hat{S}(t_{(x+1)})$ equals the probability of death occurrence at the age moment $t_{(x)}$ for $x = 1, 2, ..., I - 1$. Meantime, $p_I = \hat{S}(t_{(I)})$ is the probability of death occurrence at the last age moment $t_{(I)}$. If the last event is a death, $d_I > 0$ and $c_I = 0$, then (2.2) is replaced by

$$LE = t_{(1)} \cdot p_1 + t_{(2)} \cdot p_2 + ... + t_{(I)} \cdot p_I\,. \tag{2.3}$$

For case when the last event is a censor, $d_I = 0$ and $c_I > 0$, one can suppose assume a death event would occur very shortly after the censor. The estimator can then be suitably modified by changing status of the last event from ”censor” into ”death” and then applying (2.3). This tactic provides an underestimation as it is discussed in (Lee and Wang 2003, p. 74). However, the shortened estimation is very close to the estimation obtained when the observation process would be prolonged till the last death occurred.

Note that (2.3) is the discrete form of (2.1) when the random variable T has a discrete distribution. Therefore (2.3) yields the estimation of life expectancy as

$$LE = \sum_{x=1}^{I} t_{(x)} \cdot p_x = \sum_{x=1}^{I-1} t_{(x)} \cdot [\hat{S}(t_{(x)}) - \hat{S}(t_{(x+1)})] + t_{(I)} \cdot \hat{S}(t_{(I)})\,. \tag{2.4}$$

This can be called as the *Kaplan–Meier estimation of life expectancy*.

2.2 Kaplan–Meier estimation for semi-cohort data

Usually in annually reported demographic data, persons are not observed along the whole life time as in cohort data. In particular, due to the one calendar year duration of observation, the information is concerning to only 1 year intervals in people' life time. Therefore, the original version of the Kaplan–Meier estimation method mentioned in the previous subsection can not be applied directly to the whole population of data. For that, we can split the population into age groups of people having age in the 1 year intervals [0; 1), [1; 2), ..., [L; $L+1$), where L is the highest integer age in the data. Then apply the aforementioned Kaplan–Meier estimation method to obtain the piecewise local survival function $\hat{S}_j$ for each age interval [j; $j+1$), $j = 0, 1, ..., L$. Merging those local survival functions together, we obtain an estimation of the entire survival function for the concerned data population. Finally, the estimation of entire survival function can be combined with (2.4) to give an estimation of life expectancy.

To distinguish from the terminology "cohort data", we use the term "*semi-cohort data*" to refer a longitudinal dataset performed by following up a population during some quite short time period, instead of the whole life time of the given population. The corresponding procedure of the aforementioned Kaplan estimation when applying to the semi-semi-cohort data to obtain the life expectancy estimation is then named as *piecewise Kaplan–Meier estimation method*. The more details of the procedure is presented as follows. In each age interval [j; $j+1$), $j = 0, 1, ..., L$, let $t_1^j < t_2^j < ... < t_{I_j}^j$ be the ordered age times of deaths occurred insides the interval. Let d_x^j denote the number of deaths occurred at the age moment t_x^j and n_x^j denote the number of individuals remained in population just before the age moment t_x^j. We then can obtain the estimation of the survival function over the interval [j; $j+1$) as:

$$\hat{S}_j(t) = \prod_{t_x^j \le t} \frac{n_x^j - d_x^j}{n_x^j}\,, \tag{2.5}$$

for $t \in [j; j+1)$. Having estimation of local survival functions $\hat{S}_j(t)$ for all age intervals, we get the entire survival function as follows:

$$\hat{S}(t) = \begin{cases} \hat{S}_0(t) & for \;\; t \in [0;1)\,, \\ \left[\prod_{j=0}^{J-1} \hat{S}_j(t_{I_j}^j)\right] \times \hat{S}_J(t) & for \;\; t \in [J; J+1)\,,\; J = 1, 2, ..., L\,. \end{cases}$$

With this entire survival function, we use (2.1) to have an estimation of life expectancy

$$\begin{aligned} \mathrm{KM}_{\mathrm{LE}} = \sum_{j=1}^{L-1} \Big\{ \sum_{x=1}^{I_j-1} t_x^j \cdot [\hat{S}(t_x^j) - \hat{S}(t_{x+1}^j)] + t_{I_j}^j \cdot [\hat{S}(t_{I_j}^j) - \hat{S}(t_1^{j+1})] \Big\}, \\ + \Big\{ \sum_{x=1}^{I_L-1} t_x^L \cdot [\hat{S}(t_x^L) - \hat{S}(t_{x+1}^L)] \Big\} + t_{I_j}^L \cdot \hat{S}(t_{I_j}^L) \,. \end{aligned} \tag{2.6}$$

By leveraging as much information as possible from the concerned semi-cohort data with exact ages of all deaths and censoring, with the partition of 1 year age intervals coinciding to 1 year real observation time, the piecewise Kaplan–Meier estimation can be considered as the most accurate estimation of life expectancy.

Besides, it is evident that surveys in small areas usually provide complete mortality data with birth date and death date of every death person. Then the Kaplan–Meier estimation method can be applied to get good estimations of life expectancy for those areas. To compare the life expectancies of two areas, one can use the log - rank test (see Koletsi and Pandis 2017, for instance) to check the equality of two Kaplan–Meier survival functions. If the test is accepted, the two life expectancies can be considered to be equal. Otherwise, one can conclude the two life expectancies are different when the test is rejected and the two lines representing Kaplan–Meier survival functions on graph are separate.

3 Local parametric method of life expectancy estimation

In practice, researchers rarely have real cohort dataset, even semi-cohort dataset. Instead, routine reported dataset containing numbers pairs of deaths and of persons grouped in 5-years age groups is usually used. The widely applied Chiang method of life expectancy estimation is just based on such kind of abridged dataset, that is much more popular than cohort dataset or semi-cohort dataset. In this section we propose a novel method of life expectancy estimation for the abridged dataset with numbers of deaths and of population are aggregated along age groups.

3.1 Local parameterization of survival process

Weibull distribution was first time described in 1951 (see Weibull 1951) and is widely applied in modeling problems of various empirical research fields. Especially, Pinder III et al. (1978), as well as Juckett and Rosenberg (1993), Wilson (1994) and Bebbington et al. (2006) pointed out that survivorship data can be well fit to the Weibull distribution. For that, in this article we use the Weibull distribution to model the human survival process and to build up a new tool of life expectancy estimation.

A Weibull distribution has density function of the general form

$$f(t) = k\lambda(\lambda t)^{k-1} e^{-\lambda t^k} \quad for \ \ 0 \le t < \infty \,,$$

and is completely defined by a positive scale parameter λ and positive shape parameter k. In a survival model with the Weibull distribution, if k is less than 1, the instantaneous hazard monotonically decreases with time. If k equals 1, the instantaneous hazard is constant over time. If k is greater than 1, the instantaneous hazard increases with time. Due to this property, a Weibull distribution with a single scale parameter λ and a single shape parameter k can not be used to model the human survival time for his/her whole life span. This is because of the fact that age specific mortality rate of man decreases with time in first years of life, but stays constant over time in medium ages, and increases with time toward the end of life (see Bebbington et al. 2006, for example). Therefore this study deals with a survival model based on Weibull distribution with local parameterization in each age band of [0; 1), [1; 5), [5; 10), ..., [80; 85), and [85; ∞).

Typically, data in population studies are comprised of follow-up reports over a certain calendar year, from January 1 to December 31 of the year. Let Y be a random variable that denotes the individual age counted at the last day (December 31) of the current observation year. Then Y is used to classify the age groups [0; 1), [1; 5), [5; 10), ..., [80; 85), and [85; ∞) in abridged data. In this classification, except for the final age band, which is an open ended interval, all the early age bands are of the form $[x; x + o_x)$, with $o_0 = 1$; $o_1 = 4$ and $o_x = 5$ for all $x = 5, 10, ..., 80$. Besides, we put $o_{85} = \infty$. Let's recall that the random variable T indicates the time from birth to death of an individual. Then we can split T into the sum of random variables, each is defined on one of the age intervals [0; 1), [1; 5), ..., [80; 85), and [85; ∞), by

$$T = T \cdot \mathbf{1}_{[0;1)}(Y) + T \cdot \mathbf{1}_{[1;5)}(Y) + ... + T \cdot \mathbf{1}_{[80;85)}(Y) + T \cdot \mathbf{1}_{[85;\infty)}(Y)\,, \tag{3.1}$$

where $\mathbf{1}_{[b;c)}$ denotes the indicator function of the interval $[b; c)$,

$$\mathbf{1}_{[b;c)}(s) = \begin{cases} 1 & if \quad s \in [b; c)\,, \\ 0 & if \quad s \notin [b; c)\,. \end{cases}$$

For $x = 0, 1, 5, ..., 85$, we model the random variable $T \cdot \mathbf{1}_{[x;x+o_x)}(Y)$ to have local Weibull distribution by formulas

$$T \cdot \mathbf{1}_{[0;1)}(Y) = T_0 \cdot \mathbf{1}_{[0;1)}(Y)\,,\;\; T \cdot \mathbf{1}_{[x;x+o_x)}(Y) = (x - 1 + T_x) \cdot \mathbf{1}_{[x;x+o_x)}(Y)$$

for $x = 1, 5, ..., 80, 85$, with the random variable T_x is supposed to have the Weibull density function

$$f_x(t) = k_x\lambda_x(\lambda_x t)^{k_x-1}e^{-(\lambda_x t)^{k_x}} \quad for \;\; 0 \le t < \infty\,, \tag{3.2}$$

with cumulative distribution function given by

$$F_x(t) = \int_0^t k_x\lambda_x(\lambda_x u)^{k_x-1}e^{-(\lambda_x u)^{k_x}}du = 1 - e^{-(\lambda_x t)^{k_x}} \quad for \;\; 0 \le t < \infty\,. \tag{3.3}$$

We can see n_x is the number of individuals with value of variable Y satisfying the condition $x \le Y < x + o_x$. Besides, d_x is the number of deaths recorded in the current observation year of persons belonging to the age group $[x; x + o_x)$. It is worth to note that for the first age group $[0; 1)$, the number d_0 of deaths corresponds to the observations with the value of random variable T satisfying the condition $0 \le T \le Y < 1$, the random variable T_0 models the lifetime to death of a person born in the observation year. For $x = 1, 5, ..., 80$, the random variable T_x is defined on the age group $[x; x + o_x)$ of persons with Y as the age at December 31 satisfying the condition $x \le Y < x + o_x$. It is evident that members of this age group may die before their x-th birthday in the observation year, their age at death may take values in the interval $[x - 1; \infty)$. Therefore the number of deaths d_x, $x > 1$ corresponds to the observations of random variable T satisfying the condition $x - 1 \le Y - 1 \le T \le Y < x + o_x$. Then the random variable T_x models the remaining survival time of a member of the age group $[x; x + o_x)$, counted from his $(x - 1)$-th birthday. In the last age group $[85; \infty)$, because the death event of a person may occur before his 85-th birthday in the observation year, the age at death of an individual of this age group can take value in the interval $[84; \infty)$, the death's number d_{85} corresponds to the observations with random variable T satisfying the condition $84 \le T \le Y$. Then the random variable T_{85} models the remaining lifetime counted from his 84-th birthday.

Because an abridged dataset contains only a pair of population number n_x and deaths number d_x for each age band $[x; x + o_x)$, the data can be used only to estimate scale parameter λ_x, with a given reasonable value of the shape parameter k_x. In this study, we fix appropriate values less than 1 for k_0, k_1, k_5 and other values greater than or equal to 1 for k_x, $x = 10, 15, ..., 85$, before using the data to estimate parameters λ_x for all age groups. To facilitate for the estimation procedure of the scale parameters λ_x, we define the truncated random variables Y_0 and Y_x, $x = 1, 5, ..., 80$, by

$$Y_0 = \begin{cases} Y & if \quad 0 \le Y < 1\,, \\ 0 & if \quad Y \notin [0; 1)\,, \end{cases} \tag{3.4a}$$

$$Y_x = \begin{cases} Y - x & if \quad x \le Y < x + o_x + 1\,, \\ 0 & if \quad Y \notin [x; x + o_x + 1)\,, \end{cases} \tag{3.4b}$$

which are assumed to be uniformly distributed with the density functions

$$g_0(s) = \begin{cases} 1 & for \quad s \in [0; 1)\,, \\ 0 & otherwise. \end{cases} \tag{3.5a}$$

$$g_x(s) = \begin{cases} \frac{1}{o_x} & for \quad s \in [0; o_x + 1)\,, \\ 0 & otherwise. \end{cases} \tag{3.5b}$$

For the last age group $[85; \infty)$, the truncated random variable $Y_{85} := Y - 84$ is assumed to have an exponential distribution with a positive parameter μ and has the density function $g_{85}(s) = \mu e^{-\mu s}$ for $s \ge 0$. In the next we deal with the procedure to estimate the scale parameters λ_x, assuming the random variables T_x and Y_x are independent, $x = 0, 1, 5, ..., 85$. Note that although both T_x and Y_x depend on x, they are independent as their joint distribution density function equals to the product of the marginal distribution density functions, $f_{(T_x,Y_x)}(t,s) = f_x(t)g_x(s)$.

3.2 Estimate local scale parameter from abridged data

Typically, an abridged annual reported population dataset contains the pairs of the population number pop_x and the death number $death_x$ counted for the age group $[x; x+o_x)$, $x = 0, 1, 5, 10, ..., 85$, where $o_0 = 1, o_1 = 4, o_x = 5$ for $x = 5, 10, ..., 80$ and $a_{85} = \infty$. In this dataset, $death_x$ is the number of death events occurred at the moment when the age of the death person had been in the interval $[x; x + o_x)$. Simultaneously, pop_x is the number of alive persons at the midyear day (July 1) whose age taken at that day belonged to the same interval $[x; x + o_x)$. We refer to this kind of abridged data as midyear abridged data (MAD).

It is worth noticing that in MAD there is difference in age determination for death persons (age at death) and for alive persons (midyear age). That may affect the consistence of age-specific mortality rate calculation. Indeed, let's consider the age group [70; 75) and a person who died on April 1-st before his 70-th birthday on April 30-th. It is clear that the death must be counted for the age group [65; 70), in spite of the fact that the person should belong to the age group [70; 75) when at the midyear day of July 1-st he had completed 70 years old. Meantime, another person who died on October 30-th after his 75-th birthday in October 1-st should be of the age group [70; 75) in the midyear, whilst his death was counted for the age group [75; 80). In that situation, the numerator of any age-specific mortality rate might contain ”misplaced” deaths and lack others that had been mis-allocated to neighboring age groups. Besides, MAD might contain negative ages of new born babies with the birth day after July 1-st.

To avoid the aforementioned inconsistence, abridged data used in the LPM of life expectancy estimation are organized differently from MAD. Specifically, the last day (December 31) of the current observation year is used to get individual age. Then for each $x = 0, 1, 5, 10, ..., 85$, the number n_x is the amount of all persons whose age at the last day of year was in the interval $[x; x+o_x)$. Similarly, d_x is the number of deaths occurred in the current observation year among n_x individuals of the x-th age groups. We refer to this kind of abridged data as end-year abridged data (EAD). In practice, EAD is less popular than MAD. Therefore, it needs to recompile the number pairs $(pop_x; death_x)$ in MAD, $x = 0, 1, 5, 10, ..., 85$, to estimate the new number pairs $(n_x; d_x)$ in EDA so that the local parametric method of life expectancy estimation can be applied.

We use the Lexis diagram (Fig. 1) to describe death events in a calendar year (fixed as the 2010 year to facilitate the interpreting), related to age groups [0; 1) and [1; 5) (marked by black and gray dots). In this diagram, the horizontal axis represents the calendar time, the vertical one indicates the individual age. The horizontal coordinate of each dot informs the date of death, the vertical coordinate shows the age at death of the concerned person. Meantime, the birth date of any death person is determined by the intersection between the horizontal axis and the line parallel to the main diagonal and passing through the dot marking the death. Comparing the numbers pop_x in MAD and n_x in EAD, we can get the following observations. First, pop_0 is the number of

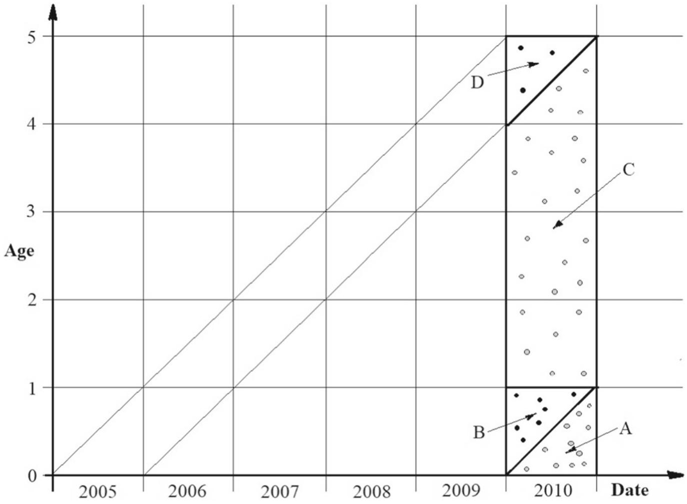

Fig. 1 Mortality pattern in the age intervals [0;1) and [1;5)

babies born during the period from July 1 of the previous year (2009) to June 30 of the currently observation year (2010). Meantime n_0 is the number of babies born between January 1 and December 31 of the currently observation year. Since there is only a half-year shift between those two periods, we can assume that the numbers of babies born in the two one-year periods (01/07/2009–30/06/2010 and 01/01/2010–31/12/2010) did not differ significantly. Thus, we can assume $n_0 = pop_0$. For the same reason, we can let $n_x = pop_x$ for all $x = 1, 5, 10, ..., 85$.

Assumed the data fit the model given by (3.1), (3.2), and (3.3), we attempt to estimate the scale parameters $\{\lambda_0, \lambda_1, \lambda_5, ..., \lambda_{80}\}$. In the Lexis diagram (Fig. 1), each dot represents one death event. Let $d_{0;1}$, $d_{0;2}$, $d_{1;1}$, and $d_{1;2}$ be the numbers of deaths occurred in the domains A, B, C, and D, respectively (by the same way we can extend the Lexis diagram and define the numbers $d_{x;1}$, $d_{x;2}$, $x = 5, 10, ..., 80$). Concerning $death_0$, the number of deaths corresponding to the dots in the both domains A and B, we see $death_0 = d_{0;1} + d_{0;2}$. It is clear that the gray dots in the domain A representing the deaths in 2010 of children born in the same year. For each of these cases, the age at death is less than the age at December 31 of 2010, which is smaller than 1. Therefore, the number $d_{0;1}$ of the such deaths is approximately equal to $n_0 \times q_{0;1}$, where due to (3.2), (3.3), (3.4a), and (3.5a), the probability of death $q_{0;1}$ is given by

$$\begin{aligned}\mathbb{P}\{0 \le T_0 \le Y_0 < 1\} &= \int_0^1 \int_t^1 f_0(t) g_0(s) ds dt = \int_0^1 (1-t) f_0(t) dt \\ &= 1 - e^{-\lambda_0^{k_0}} - \int_0^1 k_0 \lambda_0^{k_0} t^{k_0} e^{-(\lambda_0 t)^{k_0}} dt \,.\end{aligned} \tag{3.6a}$$

Whilst, the black dots in the domain B represent the deaths in 2010 of children born in 2009, each of those cases has the age at December 31 of 2009 less than the age at death, which is smaller than 1. These cases belong to the [0; 1) age group of the 2009 observation year. However, the two close years 2009 and 2010 have proximate survival models of [0; 1) age group, that implies $d_{0;2}$, the number of the deaths occurred in the domain B can be estimated by $n_0 \times q_{0;2}$, where the probability of death $q_{0;2}$ equals

$$\begin{aligned}&\mathbb{P}\{0 \le Y - 1 \le T < 1\} = \mathbb{P}\{1 \le Y_0 \le T_0 + 1 < 2\}\\ &= \int\limits_1^2 (e^{-(\lambda_0(s-1))^{k_0}} - e^{-\lambda_0^{k_0}})ds = \int\limits_0^1 e^{-(\lambda_0 t)^{k_0}}dt - e^{-\lambda_0^{k_0}} .\end{aligned} \tag{3.7a}$$

From (3.6a) and (3.7a) we get

$$death_0 \approx n_0 \times \Big\{1 - e^{-\lambda_0^{k_0}} - \int\limits_0^1 k_0 \lambda_0^{k_0} t^{k_0} e^{-(\lambda_0 t)^{k_0}} dt + \int\limits_0^1 e^{-(\lambda_0 t)^{k_0}} dt - e^{-\lambda_0^{k_0}}\Big\}.$$

Consequently, we have $\frac{k_0}{2k_0+1}\lambda_0^{2k_0} + \lambda_0^{k_0} - \frac{death_0}{n_0} \approx 0$. Then, we can estimate λ_0 by solving the quadratic equation $\alpha(\lambda_0^{k_0})^2 + \beta\lambda_0^{k_0} + \gamma = 0$ of unknown $\lambda_0^{k_0}$, with $\alpha = \frac{k_0}{2k_0+1}$, $\beta = 1$ and $\gamma = -\frac{death_0}{n_0}$. Since in the scale parameter must be positive, we take the greater solution of the quadratic equation to estimate the parameter λ_0,

$$\lambda_0 \approx \left\{\frac{-1 + \sqrt{1 + 4 \cdot \frac{k_0}{2k_0+1} \cdot \frac{death_0}{n_0}}}{\frac{2k_0}{2k_0+1}}\right\}^{1/k_0} . \tag{3.8a}$$

Consequently, we can use (3.7a) to approximate the value of $d_{0;2}$ by

$$d_{0;2} = n_0 \times \Big(\int\limits_0^1 e^{-\lambda_0^{k_0} t^{k_0}} dt - e^{-\lambda_0^{k_0}}\Big) \approx n_0 \times \Big(\int\limits_0^1 [1 - \lambda_0^{k_0} t^{k_0} + \frac{1}{2}\lambda_0^{2k_0} t^{2k_0}]dt - e^{-\lambda_0^{k_0}}\Big),$$

$$d_{0;2} \approx n_0 \times \big(1 - \frac{\lambda_0^{k_0}}{k_0 + 1} + \frac{\lambda_0^{2k_0}}{2(2k_0 + 1)} - e^{-\lambda_0^{k_0}}\big) . \tag{3.9a}$$

For the scale parameter λ_1, we deal with the deaths corresponding to the dots in the domains B, C and D. The dots in the two domains B and C represent the deaths in 2010 of the children in the age group [1; 5), that of the persons born from 2006 to 2009. Therefore, the number d_1 of the such deaths is approximately equal to $n_1 \times q_1$. Recalling (3.2), (3.3), (3.4b), and (3.5b), the probability of death q_1 can be written as

$$\begin{aligned}q_1 = \mathbb{P}\{0 \le Y_1 \le T_1 < Y_1 + 1 < 5\} &= \int\limits_0^4 \int\limits_t^{t+1} f_1(s) g_1(t) ds dt\\ &= \tfrac{1}{4}\int\limits_0^4 [e^{-(t\lambda_1)^{k_1}} - e^{-((t+1)\lambda_1)^{k_1}})]dt .\end{aligned} \tag{2.6b}$$

Simultaneously, the black dots in D represent the deaths in 2010 of persons born in 2005, those belong to the age group [1; 5) of the observation year 2009, but not to the age group [1; 5) of the observation year 2010. It is clear that the mortality models of these two age groups are close each to other. We observe that the death cases in the domain D had the age at death greater than $Y-1$ and less than 5. Additionally, recalling (3.2), (3.3), (3.4b), and (3.5b), the number $d_{1;2}$ of the deaths occurred in the domain D can be approximated by $n_1 \times q_{1;2}$, where the probability of death $q_{1;2}$ equals

$$\mathbb{P}\{4 \leq Y_1 \leq T_1 < 5\} = \int_4^5 \int_t^5 f_1(s) g_1(t) ds dt = \frac{1}{4} \int_4^5 [e^{-(t\lambda_1)^{k_1}} - e^{-(5\lambda_1)^{k_1}}] dt \,. \tag{2.7b}$$

We observe that $death_1$ is the number of deaths occurred in the domains C and D, while d_1 is the number of deaths occurred in the domains C and B. That means $death_1 = d_{1;1} + d_{1;2}$, $d_1 = d_{1;1} + d_{0;2}$ and $death_1 = d_1 + d_{1;2} - d_{0;2}$. Then (2.6b) and (2.7b) imply

$$death_1 \approx \frac{n_1}{4} \Big\{ \int_0^4 [e^{-(t\lambda_1)^{k_1}} - e^{-((t+1)\lambda_1)^{k_1}}] dt + \int_4^5 [e^{-(t\lambda_1)^{k_1}} - e^{-(5\lambda_1)^{k_1}}] dt \Big\} - d_{0;2} \,.$$

Consequently, we can use the Taylor expansion of exponential functions to have the following approximation:

$$\frac{4(death_1 + d_{0;2})}{n_1} \approx \int_0^4 [-(t\lambda_1)^{k_1} + ((t+1)\lambda_1)^{k_1}] dt + \int_4^5 [-(t\lambda_1)^{k_1} + (5\lambda_1)^{k_1}] dt \,,$$

That yields

$$\lambda_1 \approx \left\{ \frac{4(k_1+1)(death_1 + d_{0;2})}{n_1 \cdot [(k_1+1)5^{k_1} - 1]} \right\}^{1/k_1} . \tag{2.8b}$$

Then from (2.7b) we get

$$d_{1;2} \approx \frac{n_1}{4} \times \Big(\frac{-5^{k_1+1} + 4^{k_1+1}}{k_1+1} + 5^{k_1} \Big) \lambda_1^{k_1} \,. \tag{2.9b}$$

For the scale parameter λ_5, by the same argument as the above, it is clear that $death_5 = d_{5;1} + d_{5;2}$, $d_5 = d_{5;1} + d_{1;2}$ and $death_5 = d_5 + d_{5;2} - d_{1;2}$, we have the approximations

$$death_5 \approx \frac{n_5}{5} \Big\{ \int_0^5 [e^{-(t\lambda_5)^{k_5}} - e^{-((t+1)\lambda_5)^{k_5}}] dt + \int_5^6 [e^{-(t\lambda_5)^{k_5}} - e^{-(6\lambda_5)^{k_5}}] dt \Big\} - d_{1;2}$$

and

$$\lambda_5 \approx \left\{ \frac{5(k_5+1)(death_5 + d_{1;2})}{n_5 \cdot [(k_5+1)6^{k_5} - 1]} \right\}^{1/k_5} . \tag{2.8c}$$

With the estimated parameter λ_5, we approximate the value of $d_{5;2}$ by

$$d_{5;2} \approx \frac{n_5}{5} \times \left(\frac{-6^{k_5+1} + 5^{k_5+1}}{k_5+1} + 6^{k_5} \right) \lambda_5^{k_5} . \tag{2.9c}$$

We continue the above argument with the extended Lexis diagram to see $death_x = d_{x;1} + d_{x;2}, d_x = d_{x;1} + d_{x-5;2}$ and $death_x = d_x + d_{x;2} - d_{x-5;2}, x = 10, 15, ..., 80$. Then applying the approximation

$$\frac{death_x + d_{x-5;2}}{n_x} \approx \frac{\int\limits_0^5 [e^{-(t\lambda_x)^{k_x}} - e^{-((t+1)\lambda_x)^{k_x}}]dt + \int\limits_5^6 [e^{-(t\lambda_x)^{k_x}} - e^{-(6\lambda_x)^{k_x}}]dt}{5}$$

we obtain

$$\lambda_x \approx \left\{ \frac{5(k_x+1)(death_x + d_{x-1;2})}{n_x \cdot [(k_x+1)6^{k_x} - 1]} \right\}^{1/k_x} . \tag{2.8d}$$

Having the estimated parameter λ_x, we approximate the value of $d_{x;2}$ by

$$d_{x;2} \approx \frac{n_x}{5} \times \left(\frac{-6^{k_x+1} + 5^{k_x+1}}{k_x+1} + 6^{k_x} \right) \lambda_x^{k_x} . \tag{2.9d}$$

Moreover, for the open-ended age group $[85; +\infty)$, we can conclude

$$d_{85} = death_{85} + d_{80;2} . \tag{3.10}$$

3.3 Estimate life expectancy by hypothetical population

Working on a hypothetical population of $N_0 = 100000$ individuals at birth, we use the scale parameters λ_x estimated in (3.8a)–(2.8d) to calculate the probability of death q_x in each age band as follows.

$$q_0 = 1 - e^{-\lambda_0^{k_0}} , \ q_x = 1 - e^{-\lambda_x^{k_x} o_x^{k_x}} , \ x = 1, 5, 10, ..., 80 .$$

This yields the number of deaths in each age groups $D_0 = N_0 q_0$, $D_x = N_x q_x$, $x = 0, 1, 5, ..., 80$. Then the numbers of alive individuals at the beginning of each age band are completely determined as

$$N_1 = N_0 - N_0 q_0 , \ N_5 = N_1 - N_1 q_1 , \ \ldots , \ N_{85} = N_{80} - N_{80} q_{80} .$$

It is evident that for each age interval $[x; x + o_x)$, the total lifetime within the interval of those without death equals $(N_x - D_x) \times o_x$. At the same time, the density function (3.1) can be used to determine the average remaining lifetime in the age band of people with death event in the interval. In particular, for the age group [0; 1), the average remaining lifetime (denoted by DRL_0) in the age band of people with death event in the interval is equal

$$\mathbb{E}_{[0;1)}[T_0] = \frac{\int_0^1 s f_0(s) ds}{\mathbb{P}\{0 \leq T_0 < 1\}} = \frac{\int_0^1 k_0 \lambda_0^{k_0} s^{k_0} e^{-\lambda_0^{k_0} s^{k_0}} ds}{1 - e^{-\lambda_0^{k_0}}},$$

where $\mathbb{E}_{[0;1)}[T_0]$ is the expectation of random variable T_0 restricted on the interval [0; 1). In the above equation, to simplify the integration of exponential function $k_0 \lambda_0^{k_0} s^{k_0} e^{-\lambda_0^{k_0} s^{k_0}}$, we can approximate the power series by finite sum up to the 1-st order:

$$\int_0^1 k_0 \lambda_0^{k_0} s^{k_0} e^{-\lambda_0^{k_0} s^{k_0}} ds = \int_0^1 k_0 \lambda_0^{k_0} s^{k_0} \Big[\sum_{j=0}^{\infty} \frac{(-\lambda_0^{k_0} s^{k_0})^j}{j!} \Big] ds \approx \frac{k_0 \lambda_0^{k_0}}{k_0 + 1} - \frac{k_0 [\lambda_0^{k_0}]^2}{2k_0 + 1}.$$

That yields $\mathrm{DRL}_0 \approx \frac{1}{1-e^{-\lambda_0^{k_0}}} \cdot \left[\frac{k_0 \lambda_0^{k_0}}{k_0+1} - \frac{k_0 [\lambda_0^{k_0}]^2}{2k_0+1} \right]$. Then the total remaining lifetime in the age band (denoted by TRL_0) is equal to

$$(N_0 - D_0) + D_0 \cdot \mathbb{E}_{[0;1)}[T_0] = N_0 - D_0 + \frac{D_0}{1 - e^{-\lambda_0}} \cdot \left[\frac{k_0 \lambda_0^{k_0}}{k_0 + 1} - \frac{k_0 [\lambda_0^{k_0}]^2}{2k_0 + 1} \right]. \tag{3.11a}$$

For the age group $[x; x + o_x)$, $x = 1, 5, 10, ..., 80$, the average remaining lifetime in the age band of people with death in the interval is given by

$$\mathrm{DRL}_x := \mathbb{E}_{[0;o_x)}[T_x] = \frac{\int_0^{o_x} s f_x(s) ds}{\mathbb{P}\{0 \leq T_x < o_x\}} = \frac{\int_0^{o_x} k_x \lambda_x^{k_x} s^{k_x} e^{-\lambda_x^{k_x} s^{k_x}} ds}{1 - e^{-\lambda_x^{k_x} o_x^{k_x}}},$$

where $\mathbb{E}_{[0;o_x)}[T_x]$ denotes the expectation of random variable T_x restricted on the interval $[0; o_x)$. To simplify the integration of exponential function $k_x \lambda_x^{k_x} s^{k_x} e^{-\lambda_x^{k_x} s^{k_x}}$, we approximate the power series up to the 1-st order:

$$\int_0^{o_x} k_x \lambda_x^{k_x} s^{k_x} e^{-\lambda_x^{k_x} s^{k_x}} ds \approx \frac{k_x \lambda_x^{k_x} o_x^{k_x+1}}{k_x + 1} - \frac{k_x [\lambda_x^{k_x}]^2 o_x^{2k_x+1}}{2k_x + 1}.$$

Then $\mathrm{DRL}_x \approx \frac{k_x \lambda_x^{k_x}}{1-e^{-\lambda_x^{k_x} o_x^{k_x}}} \times \left\{\frac{o_x^{k_x+1}}{k_x+1} - \frac{\lambda_x^{k_x} o_x^{2k_x+1}}{2k_x+1}\right\}$. That implies the total remaining lifetime in the age band equals

$$\mathrm{TRL}_x = (N_x - D_x) \times o_x + \frac{D_x \cdot k_x \lambda_x^{k_x}}{1 - e^{-\lambda_x^{k_x} o_x^{k_x}}} \times \left\{\frac{o_x^{k_x+1}}{k_x+1} - \frac{\lambda_x^{k_x} o_x^{2k_x+1}}{2k_x+1}\right\}. \quad (3.11b)$$

For the last age interval [85; ∞), the random variable $T_{85} := T - 84$ has Weibull distribution with density function $f_{85}(t) = \lambda_{85} e^{-\lambda_{85} t}$ for $0 \leq t < \infty$. We also assume that random variable $Y_{85} := Y - 84$ has an exponential distribution with parameter μ, that means $g_{85}(t) = \mu e^{-\mu t}$ for $0 \leq t < \infty$. Then, for a person from the age group [85; ∞), the probability of death occurred in the current follow-up year is equal

$$q_{85c} = \mathbb{P}\{Y_{85} - 1 \leq T_{85} < Y_{85} < \infty\} = \int\limits_0^\infty \int\limits_{t-1}^t f_{85}(s) g_{85}(t) ds dt$$
$$= \mu[e^{\lambda_{85}} - 1] \cdot \int\limits_0^\infty e^{-(\lambda_{85}+\mu)t} dt = \frac{\mu(e^{\lambda_{85}} - 1)}{\lambda_{85} + \mu}.$$

Similarly, the probability of death occurred in future after the end of the current follow-up year, for a person of the age group [85; ∞), is equal

$$q_{85f} = \mathbb{P}\{Y_{85} \leq T_{85} < \infty\} = \int\limits_0^\infty \int\limits_t^\infty f_{85}(s) g_{85}(t) ds dt$$
$$= \mu \int\limits_0^\infty e^{-(\lambda_{85}+\mu)t} dt = \frac{\mu}{\lambda_{85} + \mu}.$$

It is clear that $q_{85c} + q_{85f} = 1$, therefore $1 = \frac{\mu(e^{\lambda_{85}}-1)}{\lambda_{85}+\mu} + \frac{\mu}{\lambda_{85}+\mu} = \frac{\mu}{\lambda_{85}+\mu} \cdot e^{\lambda_{85}}$. Thus, $\frac{\mu}{\lambda_{85}+\mu} = e^{-\lambda_{85}}$, $\lambda_{85} = -\ln(1 - p_{85c})$. Taking the ratio d_{85}/n_{85} as an estimate of q_{85c}, the last equality yields $\lambda_{85} = -\ln\left(1 - \frac{d_{85}}{n_{85}}\right) = \ln\left(\frac{n_{85}}{n_{85}-d_{85}}\right)$. Therefore, due to $T_{85} = T - 84$, the average remaining lifetime (remaining life expectancy) of persons in the age group [85; ∞) equals

$$\mathrm{RLE}_{85} = \mathbb{E}[T_{85}] - 1 = \frac{1}{\lambda_{85}} - 1 = \frac{1}{\ln\left(\frac{n_{85}}{n_{85}-d_{85}}\right)} - 1.$$

Consequently, the total remaining lifetime of all persons in the age group $[85; \infty)$ from the age of 85 years is summed as

$$\mathrm{TRL}_{85} = N_{85} \cdot \mathrm{RLE}_{85} = N_{85} \cdot \left\{\mathbb{E}[T_{85}] - 1\right\} = N_{85} \cdot \left\{\frac{1}{\ln\left(\frac{n_{85}}{n_{85}-d_{85}}\right)} - 1\right\}, \tag{3.11c}$$

where $d_{85} = death_{85} + d_{80;2}$ by virtue of (3.10).

Now summing total remaining lifetimes of age bands given in (3.11a), (3.11b) and (3.11c), dividing by the number $N_0 = 100000$ of the hypothetical population, we get the desired estimation of life expectancy:

$$\mathrm{LE} = \frac{\mathrm{TRL}_0 + \mathrm{TRL}_1 + \mathrm{TRL}_5 + \ldots + \mathrm{TRL}_{80} + \mathrm{TRL}_{85}}{N_0}.$$

The results presented above in this section can be used to create an Excel worksheet for life expectancy estimation by local parametric method.

4 Validation of life expectancy estimation methods

In this section a longitudinal survival dataset created by FilaBavi (see Nguyen and Vinod 2003) is used to verify the validity of our proposed LPM method of life expectancy estimation, comparing to Chiang method.

4.1 FilaBavi longitudinal dataset

FilaBavi, an epidemiological field laboratory sited in the Bavi District, was created by Vietnam Health Strategy and Policy Institute in 1999, with the assistance of Swedish Sida/SAREC, and conducted by Hanoi Medical University (see Nguyen and Vinod 2003). Bavi is a district in northern Vietnam, 60 km west of Hanoi. The District contains approximately 235,000 people. A random sampling of villages, with probability proportional to population size in each unit, was performed. This sample consists of 67 population clusters selected with a reported population size of 51,024 inhabitants in 11,089 households, achieved with approximately 20% of the total population of the district.

The data were collected through quarterly demographic surveillance of vital events among the study population subsequently. The quarterly surveys have been carried out at the beginning of 1999, including data on marital status changes, migrations, pregnancy follow-ups, births, and deaths. The strictly management of data collection process ensures a high quality of obtained data, allowed to track almost all vital events, including births and deaths, occurred any when in the research setting.

The dataset of FilaBavi includes first interviews in 1999, March, and last interviews in 2015, October. That means there are 15 years (2000 to 2014) of complete observation recoded in this dataset. Extracting from the dataset , this study uses the data file that

Table 1 Life expectancy estimations and residuals

Year	K-M Est	Chiang Est	Chiang Res	LPM Est	LPM Res
2000	76.02	76.33	0.31	75.56	−0.47
2001	76.03	76.30	0.28	75.54	−0.49
2002	75.65	76.09	0.44	75.15	−0.50
2003	77.43	77.47	0.04	76.54	−0.89
2004	76.58	77.26	0.68	76.57	−0.01
2005	75.69	76.08	0.39	75.73	0.03
2006	76.44	77.30	0.86	76.61	0.17
2007	77.98	78.74	0.76	78.27	0.30
2008	76.20	76.96	0.75	76.47	0.27
2009	76.88	77.15	0.27	76.86	−0.01
2010	77.05	77.61	0.56	77.43	0.38
2011	76.86	76.79	−0.06	76.89	0.03
2012	77.26	77.32	0.07	77.21	−0.05
2013	78.94	79.52	0.58	79.10	0.16
2014	80.05	80.09	0.04	80.30	0.25
Average	77.00	77.40	0.40	76.95	−0.06

has been reorganized by splitting into 15 one-year observation semi-cohort data files, each of them is related to one specific year among the 2000 to 2014 years. In the sequel the 1-year observation semi-cohort data files are used to calculate life expectancy of the Kaplan–Meier method, serving as the gold standard for validation of other life expectancy estimation methods.

4.2 Validate life expectancy estimation methods

To evaluate the accuracy of life expectancy estimation methods, we use 15 one-year semi-cohort datasets described in Sect. 4.1 to get Kaplan–Meier gold standards of life expectancy estimation determined by (2.6). In Table 1, the gold standards are given in the column ”K-M Est” representing Kaplan–Meier method life expectancy estimations of 15 years (2000 - 2014). Each 1-year semi-cohort dataset is used to make the respective MAD as entry to estimate the life expectancy by Chiang method. The estimations are recorded in the ”Chiang Est” column of Table 1. Then, differences between the Chiang estimations and the Kaplan–Meier gold standards are taken as estimation residuals in the ”Chiang Res” column of the table. From these residuals, we observe that in most of the cases, Chiang method gives over-estimated results. In particular, Chiang life expectancy estimations differ from corresponding Kaplan–Meier life expectancy estimations less than 0.3 of a year time (approximated 110 days) only at 6 among 15 years of observation. Moreover, the average of the residuals equals to 0.4 of one year time (about 146 days), that confirms the overestimation of the method on life expectancy.

The above mentioned MAD's are also used as input datasets for the excel worksheet constructed to compute life expectancy estimation according to the LPM represented in Sect. 3. To do that, the local Weibull distribution shape parameters k_x, $x = 0, 1, 5, ..., 80, 85$, must be determined in ahead and used consistently unchanged for any dataset. To find out an appropriate sequence of shape parameters' values, initially we used the dataset described in Sect. 4.1 to estimate the shape parameters, getting

$$\{k_x\} \approx \{0.151; 0.157; 0.56; 1.02; 1.35; 1.36; 0.89; 1.06; 0.97; 0.82; \\ 0.87; 0.95; 1.00; 1.01; 1.02; 0.98; 1.1; 0.97; 0.97\}. \tag{4.1}$$

However, this series applied directly to the excel worksheet does not make good estimation of life expectancy. This may be caused by the fact that, for each age group, the shape parameter in (4.1) was estimated based on the data with only mortality events occurred within the respective age interval, without death events happened in higher ages, consequently the estimate is biased. To solve the problem, we try to estimate by the LPM, but the sequence of the local shape parameters k_x in (4.1) is replaced by other combinations of local shape parameter values. One of those combinations is presented in the following series:

$$\{k_x\} = \{0.1; 0.2; 0.9; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 2; 2; 2; 1\}, \tag{4.2}$$

where $k_0 = 0.1$; $k_1 = 0.2$; $k_5 = 0.9$; $k_x = 1$ for the medium age groups with $x = 10, ..., 65$; $k_x = 2$ for three older age groups with $x = 70, 75, 80$; and $k_{85} = 1$. The column "LPM Est" of Table 1 contains the estimation results yielded by applying the LPM with local shape parameters k_x listed in (4.2). The corresponding residuals in the next by column "LPM Res" demonstrate the accuracy of the LPM, as distances from 10 estimated values to the corresponding values of Kaplan–Meier estimation are less than 0.3 of one year time. Besides, the average of the residuals counted equal 0.06 (with minus sign) of 1 year time (about 22 days), that ensures once more the exactness of this life expectancy estimation method.

5 Discussion and conclusion

The study proposed two novel methods of life expectancy estimation that are applicable to various annual reported demographic datasets. The first one, named as Kaplan–Meier estimation method, extracting complete information from data fully recorded birth date and death date of all death individuals, provided the most accurate estimation of life expectancy. Therefore, that method can be adopted as a gold standard in the accuracy investigation of other life expectancy estimations. Especially, the Kaplan–Meier estimation method can be applied to get good estimations of life expectancy for small areas, where complete mortality data can be regularly produced by routine annual surveys. The life expectancy estimates can be used to track the ongoing population health in the areas. To compare the life expectancies between two areas, or between two years, one can use the log—rank test to check the equality of two Kaplan–Meier survival functions.

The second method called as local parametric method, tailored according to the theoretical background of survival process with local parametric Weibull distributions, can be applied to abridged datasets containing only a pair of number of deaths and number of persons in each age group. The validation by using the gold standard of Kaplan–Meier estimation method showed that the local parametric method can provide very exact life expectancy estimations for 10 among 15 one-year semi-cohort datasets. Simultaneously, the validation also pointed out that the ordinary method of Chiang is an overestimation method.

However, some theoretical details should be clarified to strengthen the advantages of the local parametric estimation method. The first point is related to the series (4.2) of local shape parameters' values that were chosen somewhat heuristically. It would be interesting to verify if the parameters in (4.2) can be used as universal shape parameters applied in the here proposed method to estimate life expectancy for all other abridged datasets.

Besides, the assumption of person age evenly distributed within each age band, modeled by the uniform density functions in (3.5a) and (3.5b), can be adapted only for populations of stable age structure with population pyramid of U—inverted shape (that of the developed countries). This assumption may not be appropriate for many low - income countries with population pyramids of triangular shape. Therefore, it needs to investigate a modified version of the local parametric estimation method, that is free of that assumption.

Moreover, the local parameterization model, which contains a large number of parameters, could potentially lead to overfitting problems. Besides, when the sample size is small, the localization might lead to inefficient estimation. Therefore, advanced investigations should be proceeded to get more knowledge on those topics.

Another open problem is to determine the variance of the life expectancy estimated by the local parametric estimation method. Usually, the variance is taken to create the confidence interval of the estimate, that is necessary to use in comparison between different life expectancy values. This problem should be put for further investigations. Although the need of improvement, the proposed local parametric method of life expectancy estimation can be applied to abridged datasets instead of the ordinary method of Chiang, because of the mentioned advantages of the new method.

Acknowledgements The analysis in Sect. 4 of this study has been realized by using the longitudinal dataset of FilaBavi—Ba Vi Epidemiological Field Laboratory. Thanks are due to Prof. Dr. Nguyen Thi Kim Chuc, the main coordinator of FilaBavi, and to other scientists and workers who participated in conducting many years surveys and processing to produce the dataset, that plays a critical role for this study.

References

Bebbington M, Lai CD, Zitikis R (2006) Useful periods for lifetime distributions with bathtub shaped hazard rate functions. IEEE Trans Reliab 55(2):245–251

Chiang CL (1972) On constructing current life tables. J Am Stat Assoc 67(339):538–541

Chiang CL (1978) Life table and mortality analysis. World Health Organization, Geneva

Chiang CL (1984) The life table and its applications. Robert E Krieger Publ Co, Malabar

Cristia JP (2009) Rising mortality and life expectancy differentials by lifetime earnings in the United States. Working paper no. 665, Inter-American Development Bank, Research Department, Washington, DC

Eyler JM (1979) Victorian social medicine. The ideas and methods of William Farr. Johns Hopkins University Press, Baltimore

Eayres D, Williams ES (2004) Evaluation of methodologies for small area life expectancy estimation. J Epidemiol Commun Health 58:243–249

Fifth annual report of the registrar general, 1843. HMSO, London

Golbeck AL (1986) Probabilistic approaches to current life table estimation. Am Stat 40(3):185–190

Greville TNE (1943) Short methods of constructing abridged life tables. Rec Am Inst Actuar 32(Part 1):29–42

Griffiths C, Fitzpatrick J (2001) Geographic inequalities in life expectancy in the United Kingdom, 1995–1997. Health Stat Q 09:16–28

Hsieh JJ (1991) A general theory of life table construction and a precise abridged life table method. Biom J 33(2):143–162

Juckett DA, Rosenberg B (1993) Comparison of the Gompertz and Weibull functions as descriptors for human mortality distributions and their intersections. Mech Aging Dev 69(1–2):1–31

Kaplan EL, Meier P (1958) Non-parametric estimation from incomplete observations. J Am Stat Assoc 53:57–81

Kaplan RM, Spittel ML, Zeno TL (2014) Educational attainment and life expectancy. Policy Insights Behav Brain Sci 1(1):189–194

Keyfitz N (1966) A life table that agrees with the data. J Am Stat Assoc 61:305–311

Keyfitz N (1970) Finding probabilities from observed rates or how to make a life table. Am Stat 24(1):28–33

King G (1914) On a short method of constructing an abridged mortality table. J Inst Actuar 48:294

Koletsi D, Pandis N (2017) Survival analysis, part 2: Kaplan-Meier method and the log-rank test. Am J Orthod Dentofac Orthop 152(4):569–571

Lee ET, Wang JW (2003) Statistical methods for survival data analysis. Wiley, Hoboken

Manton KG, Stallard E, Tolley HD (1991) Limits to human life expectancy: evidence, prospects, and implications. Popul Dev Rev 12(4):603–637

Meulen A (2012) Life tables and survival analysis. Statistics Netherlands, The Hague/Heerlen

Nguyen TKC, Vinod KD (2003) FilaBavi, a demographic surveillance site, an epidemiological field laboratory in Vietnam. Scand J. Public Health 31(Suppl. 62):3–7

OECD (2017) Life expectancy, in pensions at a glance 2017: OECD and G20 indicators. OECD Publishing, Paris

Pinder JE III, Wiener JG, Smith MH (1978) The Weibull distribution: a new method of summarizing survivorship data. Ecology 59(1):175–179

Pollard JH (1989) On the derivation of a full life table from mortality data recorded in five-year age groups. Math Popul Stud 2(1):1–14

Reed IL, Merrell M (1932) On a short method of constructing an abridged life table. Am J Hyg 30:33–62

Silcocks PBS (2004) Improving estimation of the variance of expectation of life for small populations. J Epidemiol Commun Health 58:611–612

Silcocks PBS, Jenner DA, Reza R (2001) Life expectancy as a summary of mortality in a population: statistical considerations and suitability for use by health authorities. J Epidemiol Commun Health 55:38–43

Toson B, Baker A (2003) Life expectancy at birth: methodological options for small populations. National Statistics Methodological Series No. 33. A National Statistics publication, Her Majestyas Stationery Office, Norwich

Weibull W (1951) A statistical distribution function of wide applicability. J Appl Mech 18(3):293–297

Wilmoth JR, Andreev K, Jdanov D, Glei DA (2007) Methods protocol for the human mortality database. University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock. http://www.mortality.org/Public/Docs/MethodsProtocol.pdf. Accessed 26 Jan 2021

Wilson DL (1994) The analysis of survival (mortality) data: fitting Gompertz, Weibull, and logistic functions. Mech Ageing Dev 74:15–33