A hazards approach to the biometric analysis of infant mortality

Göran Broström and Tommy Bengtsson

2023-03-24 15:41:26

Abstract

A variation of the Bourgeois-Pichat biometric analysis of infant mortality is suggested. In the original model, cumulative mortality in the last eleven months of infancy is assumed to follow a uniform distribution given a log-cube transformation of age. Instead, we assume an exponential distribution. The practical difference is that while the denominators in the estimation are constant in the Bourgeois-Pichat model, equal to the number of births, in our model, the denominators are the sizes of the current populations at risk. We argue that our assumption is more satisfactory from a theoretical point of view, since it focus on the *conditional* probability of dying. Our model gives slightly different estimates of endogenous and exogenous mortality and, in addition, the model fit is often better, especially in cases with higher levels of infant mortality. We use data from northern and southern Sweden for the period 1710 to 1950 to compare the two methods.

Contents

1 Introduction

Jean Bourgeois-Pichat saw the first year of life not only as the period when mortality is highest, but also the period when improvements should be possible (Bourgeois-Pichat, 1951a,b; Bourgeouis-Pichat, 1952). He argued that infant deaths should be divided into two categories, endogenous and exogenous deaths. Endogenous deaths are, by definition, due to inherited factors, or acquired during gestation or delivery. They typically occur at the initial period of life, though sometimes much later and include congenital debility, prematurity, malformations, and disease in early life. Exogenous infant mortality, which Bourgeois-Pichat regards as accidental, are deaths for which the society must hold itself responsible and to reduce endogenous mortality, medical intervention is essential (Bourgeouis-Pichat, 1952). Since it is, based on causes of deaths, difficult to distinguish between endogenous and exogenous factors, either because they are inaccurate or difficult to make due to the diseases themselves, or the information do not exist, Bourgeois-Pichat offered a solution to the problem.

Bourgeois-Pichat's biometric model to differentiate between endogenous and exogenous mortality is based on an idea of a universal law governing the distribution of deaths in the first year of life (Bourgeouis-Pichat, 1952). He assumes that all deaths taking place in the last eleven months of the first year are exogenous. He argues that, although there are also some endogenous mortalities during this period, they do not affect the development of mortality much. Supported by examples from mid-twentieth century Western and Southern Europe, the US and Canada, New Zeeland, and other countries, Bourgeois-Pichat argued that the cumulative infant deaths after the first month follows a linear development given a log-cube transformation of age (Bourgeouis-Pichat, 1952). Mortality in the first month of life is then divided into an endogenous and an exogenous assuming that the same linearity also exists in this period. It leads to the conclusion that exogenous deaths in the first 28 days of life represent 22.8 percent of the deaths from the 29th to the 365th day (Bourgeouis-Pichat, 1952; Pressat et al., 1972), or 18.6 percent of the total number of exogenous infant deaths.

Bourgeois-Pichat also found deviations from the linear death pattern after the first month, the curve for Sardinia in 1948 bend strongly upward after four months (Bourgeouis-Pichat, 1952, Figure 12). Knodel and Kintner (1977) give ample examples of upward bending curves and argue that they are due to increasing mortality after stopping breast feeding. In Quebec 1944–1947, the curve bends down quite strongly after the sixth month (Bourgeouis-Pichat,

1952, Figure 8). A downward bend is also what has been found using historical Swedish parish data (Bengtsson, 1999, Figure 6b; Bengtsson and Lundh, 1999, Figures 6 and 7; Lynch et al., 1998, Figures 1 and 2; Sundin and Tedebrand, 1981, Figure 6d), though less pronounced than in Quebec. The curve for Sweden 1910–1946 show, however, no such downward bend (Bourgeois-Pichat, 1951b, Figure 9; Bourgeouis-Pichat, 1952).

Bourgeois-Pichat's biometric method has also been used to identify problems with data recording (Wrigley, 1977; Wrigley and Schofield, 1981). While a high level of endogenous mortality cannot be used as a criterion for high data quality, the opposite holds true. A very low level of endogenous mortality indeed indicates data problems, just like large deviations from the normal sex ratio at birth do. These two criteria are, in fact, often used in historical studies to evaluate data quality (Bengtsson et al., 2004). They become even more precise if they are applied to different social strata, since under-recording of early deaths often have a social gradient, possibly due to the costs involved in a burial (Bengtsson, 1999; Bengtsson and Lundh, 1999).

We suggest a variation of the Bourgeois-Pichat biometric analysis of infant mortality. Instead of assuming that the cumulative mortality in the last eleven months of infancy follows a uniform distribution, given a log-cube transformation of age, we assume an exponential distribution. The difference is that while the denominator is constant in the Bourgeois-Pichat model, equal to the number of births, in our model, the denominator is the current population at risk. This assumption is more satisfactory from a theoretical point of view, since it models the conditional probability of dying. The growth in birth weight also follows this distribution (Bourgeois-Pichat, 1951a,b). In addition, the model fit is slightly better for our model, especially in cases with high level of infant mortality, common in pre-modern societies. The advantage is that not only is our assumption more attractive from a theoretical point of view, and therefore easier to interpret, but also that it is makes it easier to estimate exogenous and endogenous infant mortality with standard survival analysis programs. We give examples of the two methods using data for 19th and 10th century norhern Sweden, where both mehods fit very well. In addition, we give examples from eighteenth Sweden, where the curve is concave during the last eleven month of infancy, despite high levels of mortality in the first month of life.

2 The Bourgeois-Pichat procedure

Central in the procedure suggested by Bourgeois-Pichat (1951a) is *the log-cube transform*, see Equation (1).

$$g(t) = C \log^3(t+1),\ 0 \leq t \leq 365, \tag{1}$$

where t is age measured in days and C is a normalizing constant,

$$C = \frac{365}{\log^3(365+1)}.$$

[Figure 1 about here.]

Table 1: Life table for infants in Västerbotten, 1861-1890.

Day	g(Day)	Population	Deaths
0.25	0.0	50034	564
1.00	0.6	49470	138
2.00	2.4	49332	81
3.00	4.7	49251	83
4.00	7.4	49168	64
361.00	363.0	44653	8
362.00	363.5	44644	6
363.00	364.0	44637	5
364.00	364.5	44632	6
365.00	365.0	44626	6

The constant C is chosen so that $g(365) = 365$, see Figure 1. Note that C is *not* part of the original definition of the log-cube transform, but provided here only to make graphical comparisons easier to interpret. It makes no difference otherwise.

Assume that a cohort of infants is followed over time from birth to age one. There are no drop-outs (no right censoring except at age 365 days). The exact age in days at each observed death is noted and transformed by g, and due to imperfect time measuring there are often tied death ages. As an illustrative example, we use a data set from northern Sweden, covering the years 1861–1950.

For the first and last 30 years in this data set, we have the results shown in Figure 2.

[Figure 2 about here.]

The cumulative numbers of deaths, divided by *the number of births*, are plotted against the corresponding death ages on the g time scale, and as is seen, the fit to linearity after 28 days (68 on the g scale) is excellent for both time periods.

3 The hazard-based procedure

Instead of using the total number of births throughout in the denominator as in the Bourgeois-Pichat procedure, we suggest using the present risk-set size, that is, the the total number of infants still alive immediately prior to the death age under investigation. Given proper data, it is easily achieved. For instance, by using the function `risksets` in the **R** (R Core Team, 2022) package `eha` (Broström, 2022, 2021) we get Table 1.

It starts with 50034 live births, of which 564 dies on the day of birth, on average at the age of six hours (a quarter of a day), and so on.

The hazard-based plot corresponding to Figure 2 is shown in Figure 3.

[Figure 3 about here.]

Both methods show an excellent fit to linearity for this data set. But in order to get a close comparison between the Bourgeois-Pichat and the hazards method, we show the hazards method in the style of Figure 2, see Figure 4.

[Figure 4 about here.]

The left hand panel is clearly more informative, we see both the exogenous and the endogenous cumulative plots. The dashed blue line in the left hand panel is the exogenous cumulative hazard function, while it in the right hand panel is the same function, but shifted upwards 0.027 units. It is a consequence of the "paper, pencil and ruler" method used by **?**, and it can be argued that it gives a visual perception of goodness-of-fit lacking in the left hand panel. But that is not relevant, because the same perception of goodness-of fit is achieved by looking at the estimated endogenous cumulative hazard function: It should be a horizontal straight line in the post-neonatal age period.

So in the following graphs, we stick to the hazards plot type.

4 Theoretical considerations

We note that the Bourgeois-Pichat method requires that no right censorings (infants lost to follow-up before one year of age) are present, and no left truncations (infants under observation only from an age later than birth). The hazards approach, on the other hand, allows left truncation and right censoring. Often, though, this not very important, since new-born babies tend to be geographically stationary for their first year of life.

4.1 Post-neonatal mortality

We start with the *postneonatal* period, since that is the period where Bourgeois-Pichat claims that the *cumulative distribution function* (CDF) is *Uniform*, and we suggest that the CDF is *exponential*, that is, the *cumulative hazard function* is "uniform" (linear).

In order to investigate this, the data set from above is *left truncated* at age 28 days (at 68 on the "g" scale), that is, we are considering the *conditional* survival distribution, given survival to age 28.

[Figure 5 about here.]

The *Exponential* fit is excellent, with a slight edge for the early period data, where post-neonatal mortality is high.

The fit of the *Uniform* distribution is already shown in Figure 2. Also a very good fits in the post-neonatal period.

Table 2: Distribution over age periods of exogenous mortality (per 1000).

Neonatal	**Total**	**Fraction**
0.930	5	18.60
1.864	10	18.64
9.478	50	18.96
19.371	100	19.37
40.581	200	20.29

4.2 Exogeneous mortality

We now consider the whole infant age span (on the g scale), and compare the uniform and exponential densities, especially of interest is the fraction of deaths that occur in the neonatal phase. Manfredini (2004) argues that a less satisfactory property of the B-P model is that this fraction is constant, 18.56 percent, independent of the overall level of infant mortality p. This is not the case for the hazards model, see Table 2. As the total mortality increases, the fraction exogenous deaths in the neonatal period also increases.

4.3 Endogeneous mortality

In order to calculate the *endogeneous* infant mortality, a simple two-step procedure leads to the goal.

1. Estimate the post-neonatal mortality following the exponential route. It reduces to a simple occurrence/exposure calculation: The total number of post-neonatal deaths D is divided by the total postneonatal exposure E on the g scale. So

   $$\hat{\lambda} = \frac{D}{E}$$

   is the estimated hazard function (constant), and the cumulative hazard function is

   $$\hat{H}(t) = \hat{\lambda}t,\ t > 0$$

2. On the full infant interval, estimate the total cumulative hazard rate $A(t)$ with the usual Nelson-Aalen estimator $\hat{A}(t)$ (Nelson, 1972; Aalen, 1978). Then, subtract $H(t)$ from $\hat{A}(t)$ to get $\hat{E}(t)$.

   $$\hat{E}(t) = \hat{A}(t) - \hat{\lambda}t,\ 0 < t \leq g(28). \tag{2}$$

   Note that subtraction and addition of competing risks are okay on the hazards scale, but not with probabilities, which is yet another argument in favor of the hazards approach.

Let us do it with the given data (Table 3), and the full period 1861–1950.

Table 3: First and last five rows of data frame.

birthdate	sex	enter	exit	event	period
1861-01-01	girl	0	163	1	1861-1890
1861-01-01	boy	0	365	0	1861-1890
1861-01-01	girl	0	365	0	1861-1890
1861-01-01	boy	28	365	0	1861-1890
1861-01-01	boy	0	57	1	1861-1890
1950-12-31	girl	0	365	0	1921-1950
1950-12-31	girl	0	365	0	1921-1950
1950-12-31	boy	0	365	0	1921-1950
1950-12-31	girl	0	365	0	1921-1950
1950-12-31	girl	0	309	0	1921-1950

4.3.1 Step 1: Estimate post-neonatal (exogenous) mortality

Post-neonatal mortality is the same as exogenous mortality in the sense that the hazard functions are the same on the post-neonatal age interval. So the first step involves data left truncated at age 28 days. The maximum-likelihood estimator of the rate in the exponential distribution is simply the number of observed post-neonatal deaths 10136, divided by the total post-neonatal exposure time 61492634 (still on the g scale), giving a rate of 1.6483275×10^{-4}.

The rate is a very small number as a consequence of the very small time unit (implying large total exposure time), originally *day*.

So we are done with the simple post-neonatal period.

4.3.2 Step 2: Estimate total and endogenous mortality

The estimation of the cumulative hazard function for the total neonatal mortality is standard.

Next, take the difference between "Total" and "Exogenous", as in Equation (2) to get "Endogenous", the result is shown in Figure 6.

[Figure 6 about here.]

4.4 The assumptions about exogenous mortality

Exogenous causes of death are due to social, cultural, economic, and environmental factors, and also accidents. An argument for the constant hazards and the uniform approaches to exogenous infant mortality is that, after compensating for age via the log-cube transformation, these factors are random and evenly spread in time, like *white noise*. However, this disregards seasonal variations in weather, and following that the spread of infectious diseases.

To illustrate this, data will be split in two ways, first by "birth month" (cohort data) and second by "calendar month" (period data).

4.4.1 Cohorts by birth month

[Figure 7 about here.]

We see that the post-neonatal part of the total cumulative hazards are more or less convex (bend upward at the end) for those born in spring (red) or summer (green), while the curves are concave for births in the winter (blue). Bad fits to "constant hazard" and "uniform distribution". However, as the solid black curve shows, the 12-month mixture has an excellent fit to a straight line.

Note the two outliers, February and October. Born in February means a tough first month of life, but surviving that, the future prospects are great. The opposite holds for those born in October, no problems during the neonatal phase, but then comes the tough winter months.

4.4.2 Periods by month

[Figure 8 about here.]

The periods are of course more homogeneous regarding weather and external threats, and it shows off in that for all months the suggested models fit reasonably well, although on different levels, here shown by the "constant hazards model" for exogenous mortality.

4.4.3 Conclusion

The assumptions about constant hazard and uniform distribution of exogenous mortality are only applicable on the population level, it can never hold for an individual. It can be thought of as an average over birth cohorts, where it is important to have birth dates spread evenly within calendar years.

4.5 Tabular data

Often the Bourgeois-Pichat method is applied to tabular data. The most extreme case is that the only available data are number of births B, number of neonatal deaths D_1, and total number of deaths D. The Bourgeois-Pichat line fitting is very simple to do in this case, but how about the hazards approach?

It turns out that it is easy to find given the data at hand followíng the procedure given in Section 4.3: The probability of dying in the post-neonatal period, given survival of the neonatal period is estimated by $\hat{p} = (D - D_1)/(B - D_1)$, and the length of the post-neonatal time span is $T = g(365) - g(28) \approx 297$ "g-units". The estimate of the rate λ in the post-neonatal period is then given by

$$\hat{\lambda} = -\frac{log(1 - \hat{p})}{T}.$$

5 Biometric analysis in practice

The original procedure of Bourgeos-Pichat is compared to the hazard based procedure for some typical cases from the real world.

5.1 Västerbotten

Thee data set from Västerbotten 1861–1950, kept by CEDAR, Umeå University (Demographic Data Base, 2022), shows almost perfect fits to both the hazards model and the original biometric model (Bourgeois-Pichat, 1951a,b) in all the three subperiods, see the results in the following three subsections.

5.1.1 The period 1921–1950

The crude IMR in Västerbotten 1921–1950 was around 56 per thousand, a rather low figure in context. Let us perform the biometric analysis with these data, see Figure 9.

[Figure 9 about here.]

There is a good hazards model fit, and we see that the endogenous mortality clearly dominates the early days of neonatal mortality, and almost vanishes towards the start of the post-neonatal period.

The B-P model fit is slightly worse, but not much to bother about.

5.1.2 The period 1890–1920

The crude IMR in Västerbotten 1891–1920 was around 95 per thousand, clearly higher than the later time period. Let us perform the biometric analysis with these data, see Figure 10.

[Figure 10 about here.]

The dominance of endogenous mortality in early life is still clear. The hazards fit is still slightly better.

5.1.3 The period 1861–1890

The crude IMR in Västerbotten 1861–1890 was around 105 per thousand, highest of the three time periods. The biometric analysis is shown in Figure 11.

[Figure 11 about here.]

The conclusion here is almost the same as for the later time periods, good model fit and endogenous dominance in the very early days of life. However, the B-P method seems to have a slight upper hand regarding model fit.

5.2 Scania 1710–1800.

The Scanian dataset covers five parishes in Scania 1710–1800 (Bengtsson et al., 2014). It is an interesting case because it appears as if the data quality is low since the linearity during the post-neonatal period is gone. Rather, there is a "hockey-stick" pattern, also seen in other old data sets. See Figure 12.

However, this data set is known to be of very high quality (Bengtsson and Lundh, 1999). Sex ratio at birth is as expected and the teonatal mortality is high for all social groups. We are therefore inclined to think that the biometric model as it is described here simply is inappropriate during the eighteenth century. Specifically, the log-cube transformation, which fits well with 19th century data, seems to be completely out of order in the eighteenth century.

[Figure 12 about here.]

5.3 Skellefteå 1820–1835

An example of severe under-registration during the neonatal period, see Figure 13. There is no room for endogenous deaths at all, but otherwise a reasonably good fit with both methods in the post-neonatal period..

[Figure 13 about here.]

6 Conclusion

If anything, the hazards method never performs worse than the B-P method. But the real strength of the hazards method is that it fits naturally into general modern survival analysis with censored and truncated data, and also proportional hazards models with covariates.

On the other hand, the whole business with the biometric analysis and the log-cube transform seems to be highly questionable. It works for data in approximately the same time period and similar environment, but beyond that, it is not generalizable at all. These methods should be used with great care, if at all.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6:701–726.

Bengtsson, T. (1999). The vulnerable child. Economic insecurity and child mortality in pre-industrial Sweden: A case study of Västanfors, 1757–1850. *European Journal of Population*, 15:117–151.

Bengtsson, T., Campbell, C., and Lee, J. Z. (2004). *Life under Pressure: Mortality and Living Standards in Europe and Asia, 1700–1900.* MIT Press.

Bengtsson, T., Dribe, M., Quaranta, L., and Svensson, P. (2014). The Scanian economic demographic database, version 4.0 (machine-readable database). Centre for Economic Demography, Lund University.

Bengtsson, T. and Lundh, C. (1999). Child and infant mortality in the nordic countries prior to 1900. Technical Report 66, Department of Economic History, Lund University, Lund. Lund Papers in Economic History.

Bourgeois-Pichat, J. (1951a). La mesure de la mortalité infantile. I. Principes et métodes. *Population (French Edition)*, 6:233–248.

Bourgeois-Pichat, J. (1951b). La mesure de la mortalité infantile. II. Les causes de décès. *Population (French Edition)*, 6:459–480.

Bourgeouis-Pichat, J. (1952). An analysis of infant mortality. Population Bulletin 2.

Broström, G. (2021). *Event History Analysis with R, Second Edition*. Chapman & Hall/CRC, Boca Raton.

Broström, G. (2022). *eha: Event History Analysis*. R package version 2.10.1. https://CRAN.R-project.org/package=eha.

Demographic Data Base (2022). U22003. doi: https://dx.doi.org/10.17197/U22003.

Knodel, J. and Kintner, H. (1977). The impact of breast feeding patterns on the biometric analysis of infant mortality. *Demography*, 14:391–409.

Lynch, K. A., Greenhouse, J. B., and Brändström, A. (1998). Biometric modeling in the study of infant mortality. *Historical Methods*, 31(2):53–64.

Manfredini, M. (2004). The bourgeois-pichat's method and the influence of climate: New evidence from late 19th-century Italy. *Social Biology*, 51:24–36.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14:945—-965.

Pressat, R., Matras, J., and Keyfitz, N. (1972). *Demographic Analysis; Methods, Results, Applications*. Aldine-Atherton, New York.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sundin, J. and Tedebrand, L.-G. (1981). Mortaility and morbidity in Swedish iron foundries, 1750-1875. In Brändström, A. and Sundin, J., editors, *Tradition and Transition: Studies in Microdemography and Social Change*. Demographic Data Base, Umeå University, Umeå.

Wrigley, E. (1977). Birth and baptisms: The use of anglican baptism registers as a source of information about the numbers of births in England before the beginning of civil registration. *Population Studies*, 31:281–312.

Wrigley, E. and Schofield, R. (1981). *The population history of England 1541–1871: A reconstruction*. Edward Arnold, London.

List of Figures

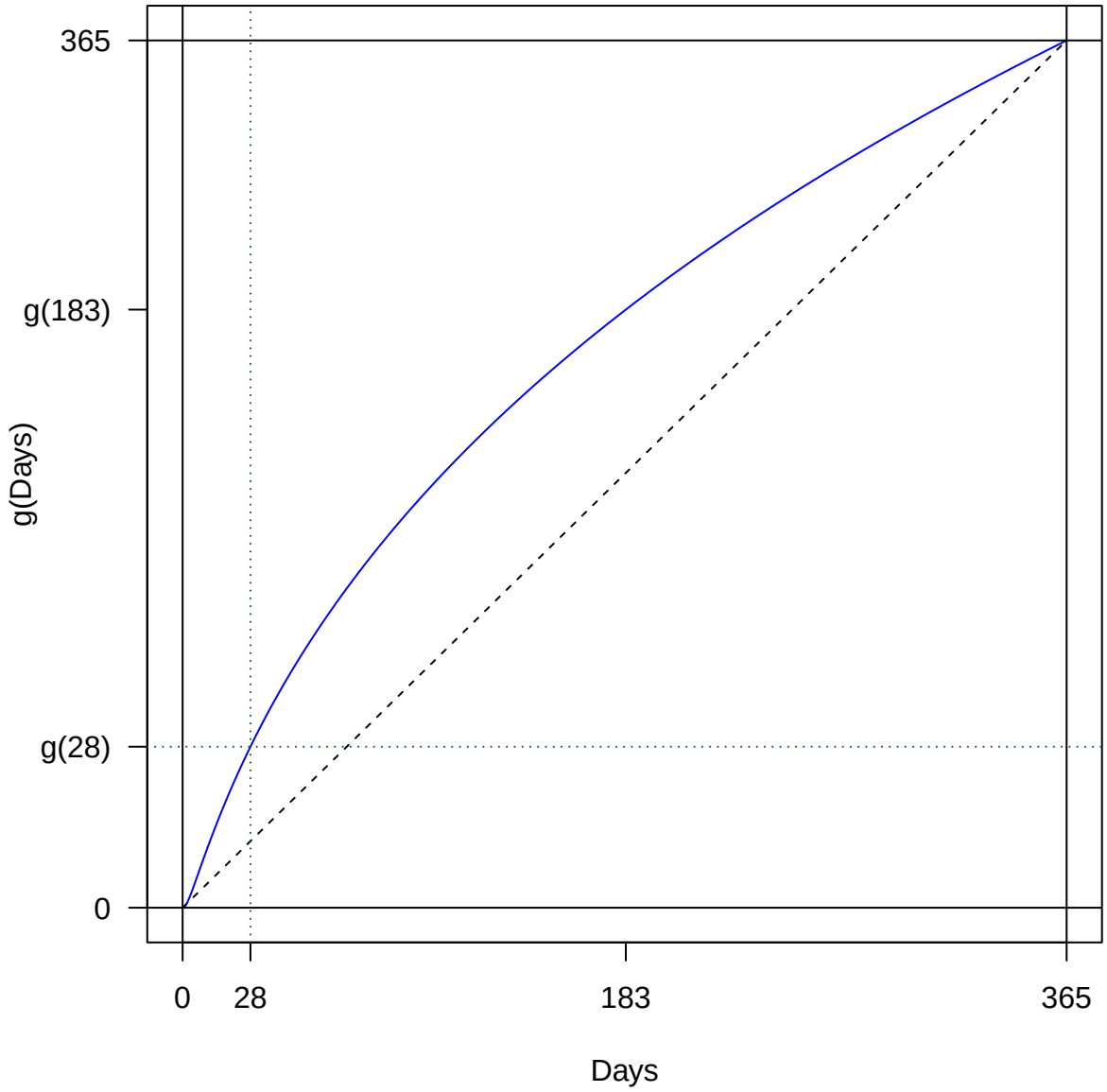

Figure 1: The log-cube transform of time in days versus the identity transform (dashed). Note that g(365) = 365.

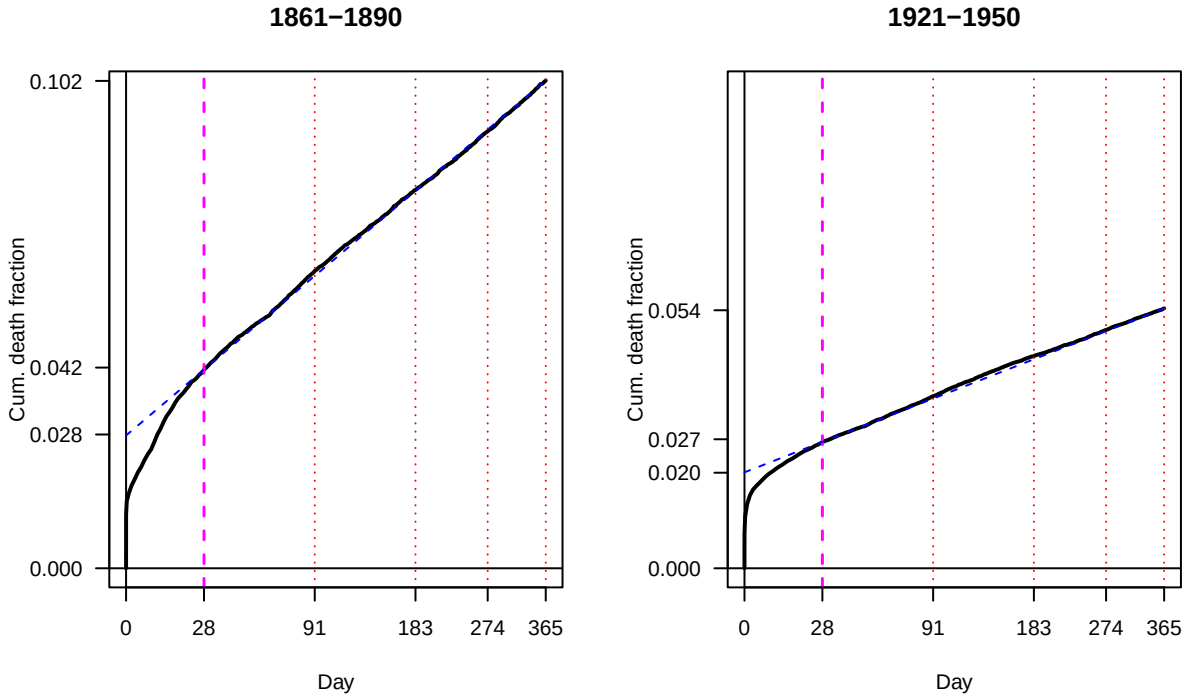

Figure 2: The periods 1861–1890 and 1921–1950 in Västerbotten, Bourgeois-Pichat method.

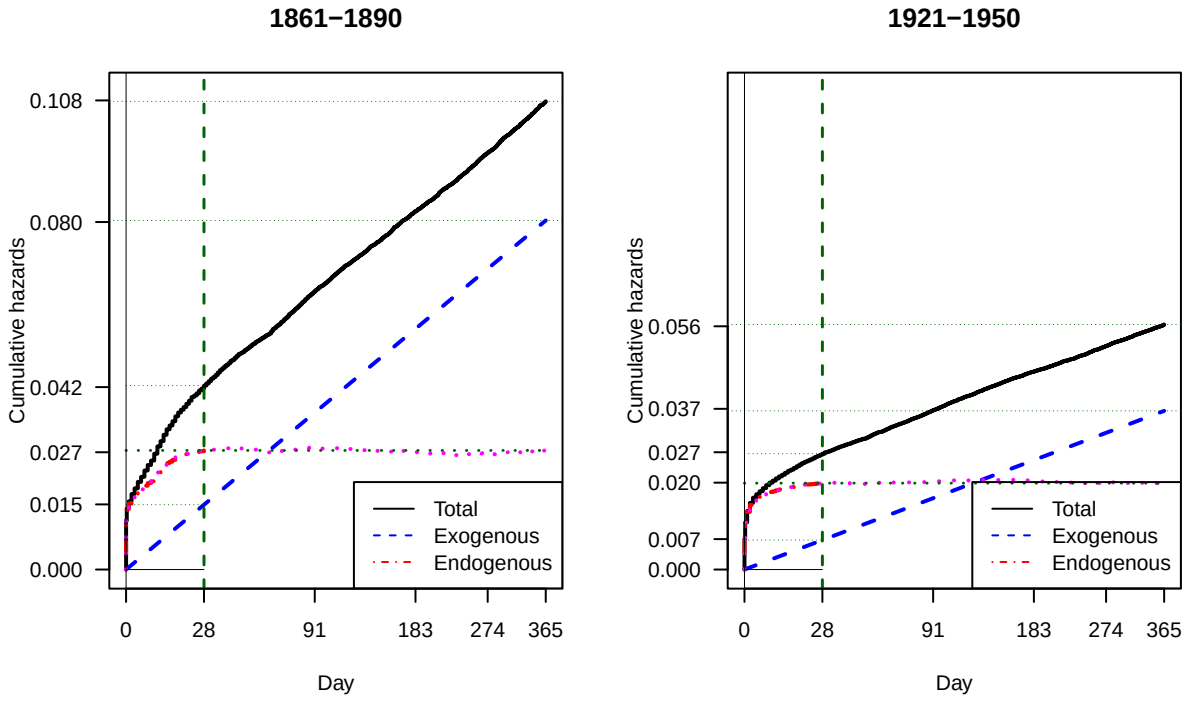

Figure 3: The periods 1861–1890 and 1921–1950 in Västerbotten, hazards method.

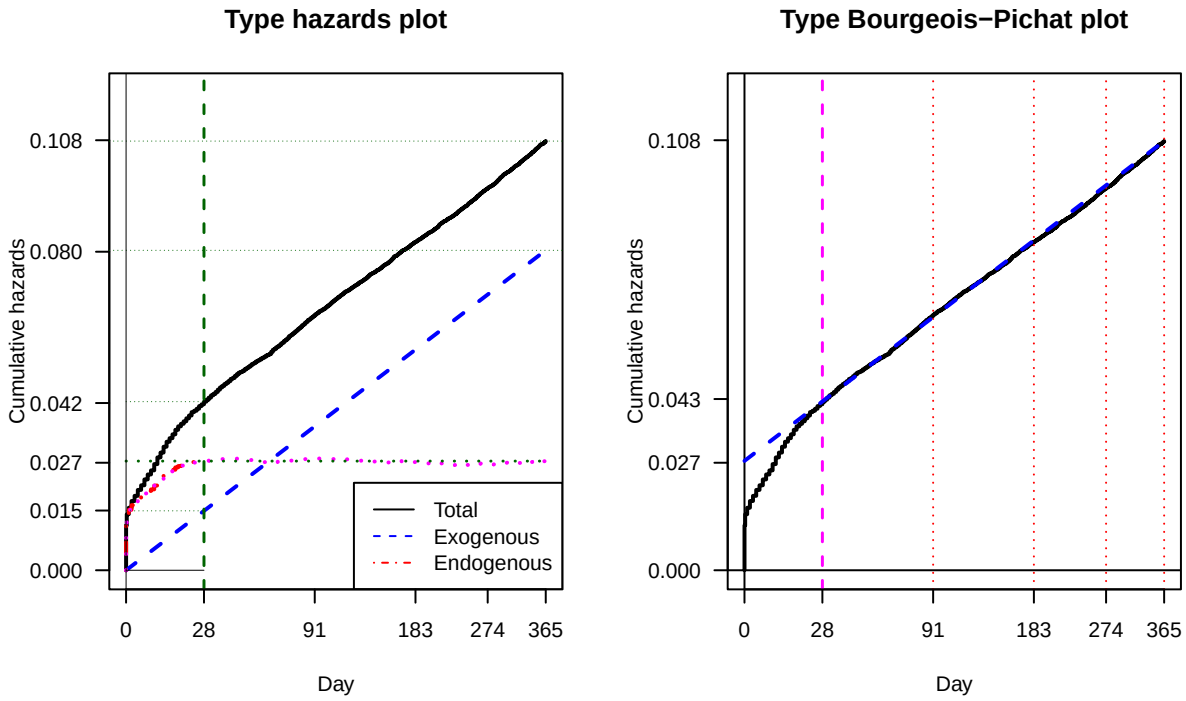

Figure 4: The hazards method plotted in the Bourgeous-Pichat way.

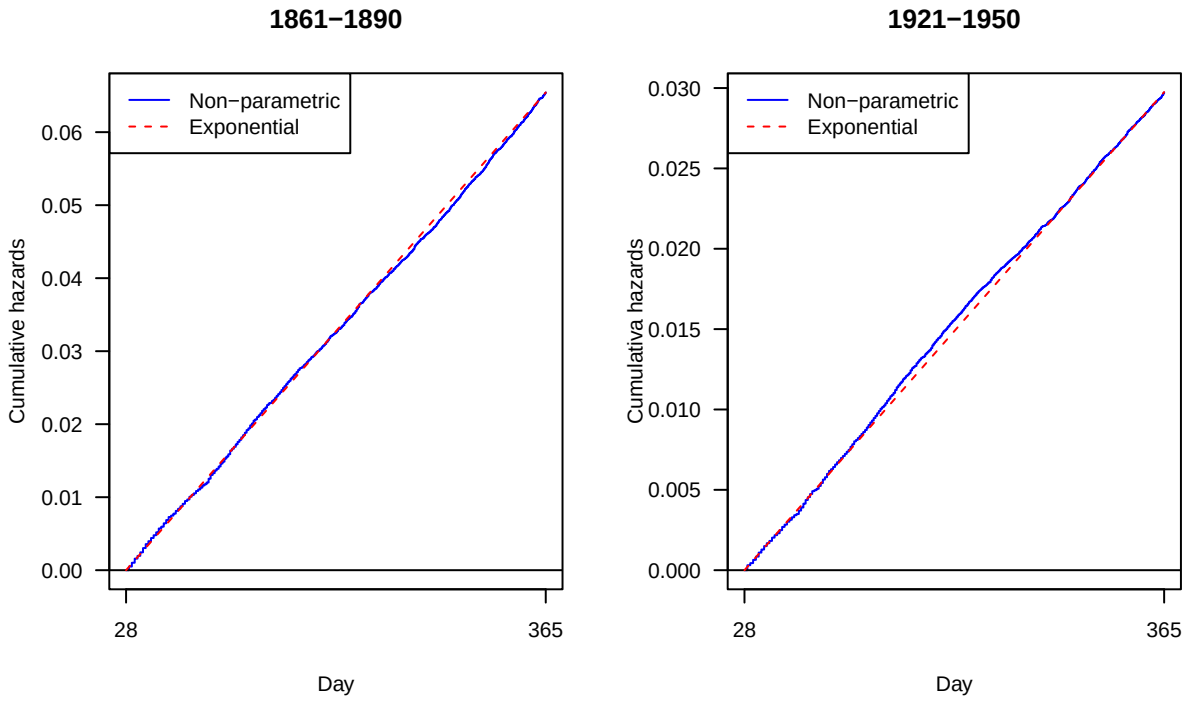

Figure 5: Exponential fits to postneonatal data, Västerbotten.

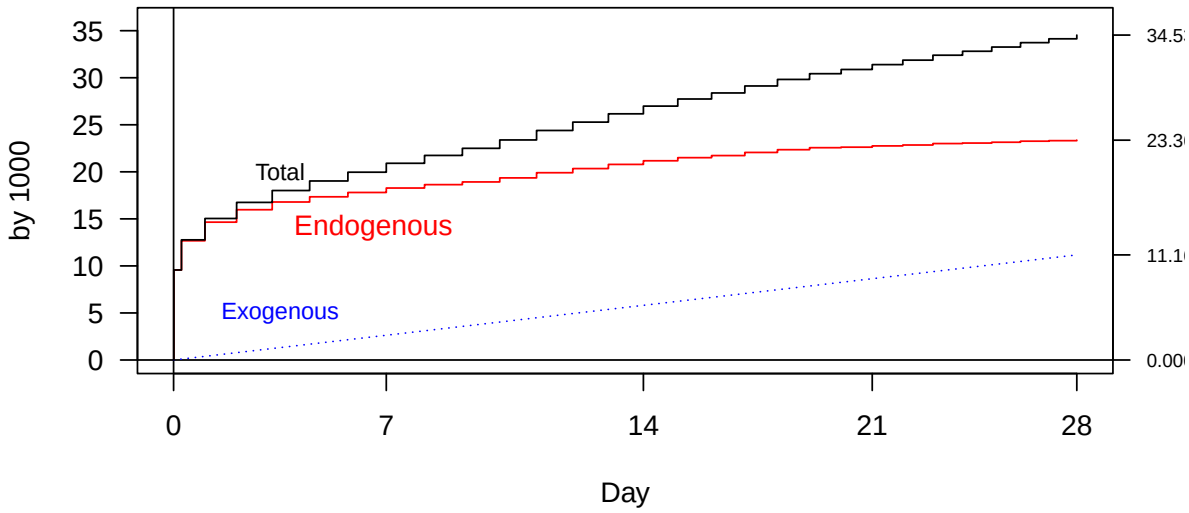

Figure 6: Cumulative hazard functions for neonatal mortality, Västerbotten.

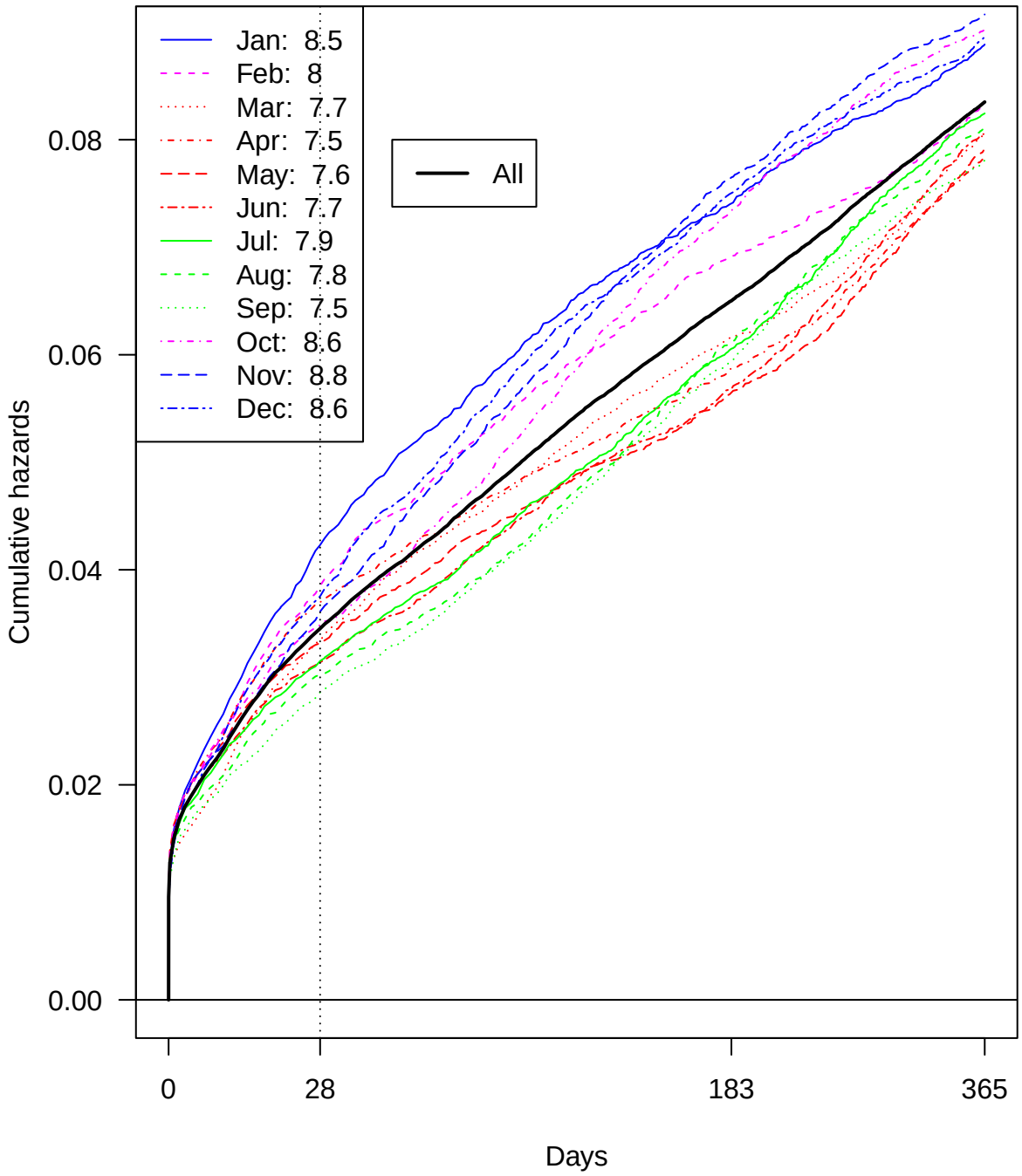

Figure 7: By birth month, cohorts, Västerbotten 1861-1950.

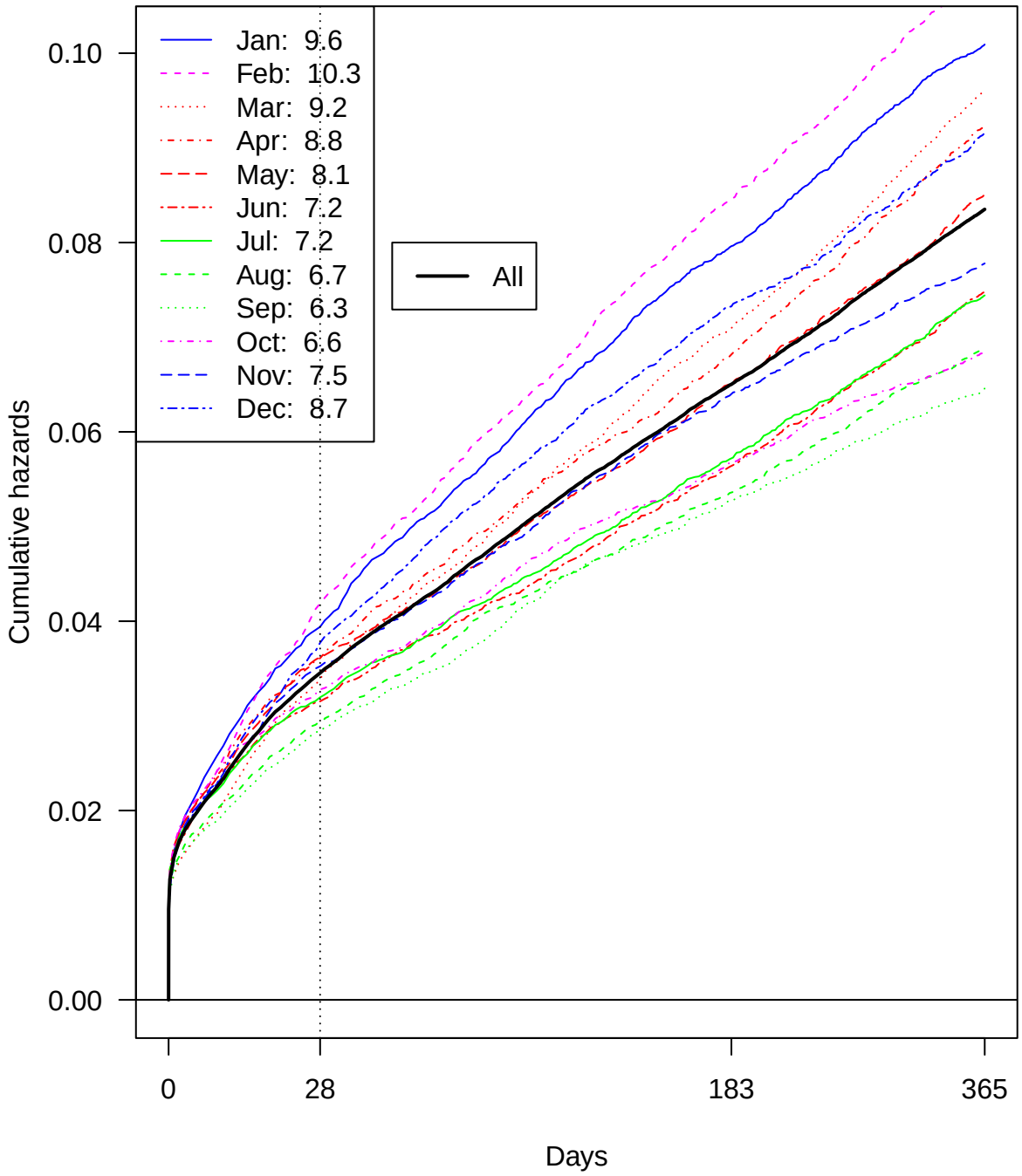

Figure 8: By month, period data, Västerbotten 1861-1950.

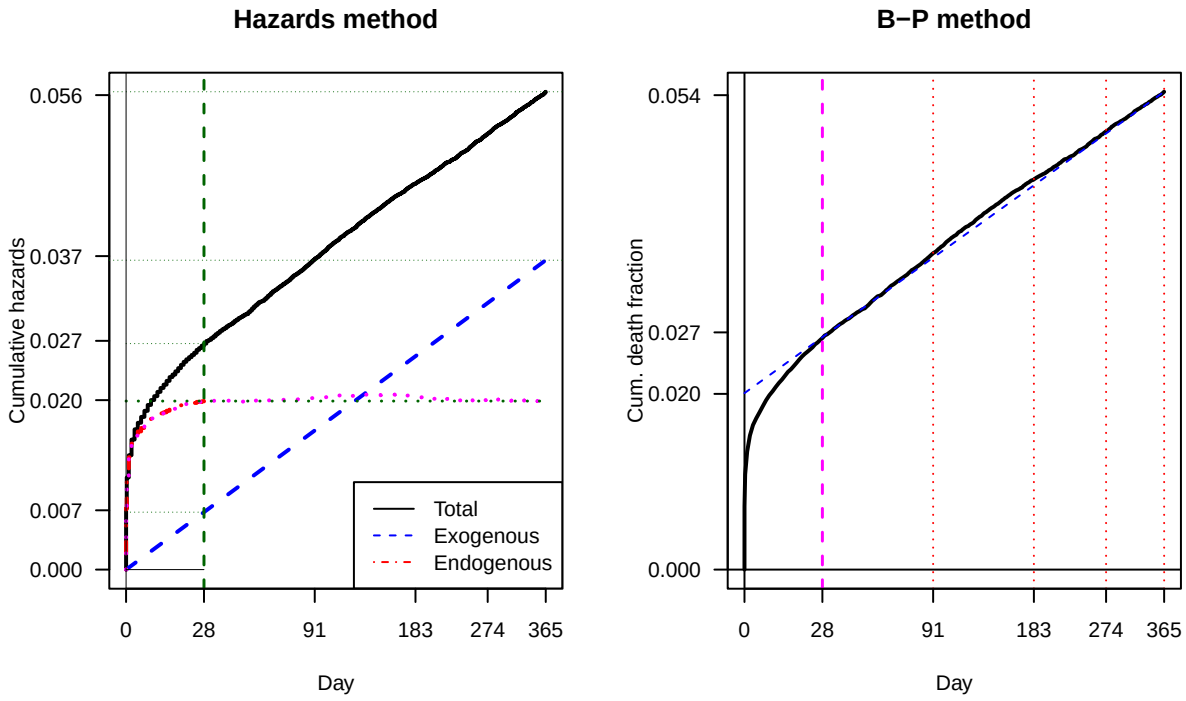

Figure 9: Västerbotten 1921–1950.

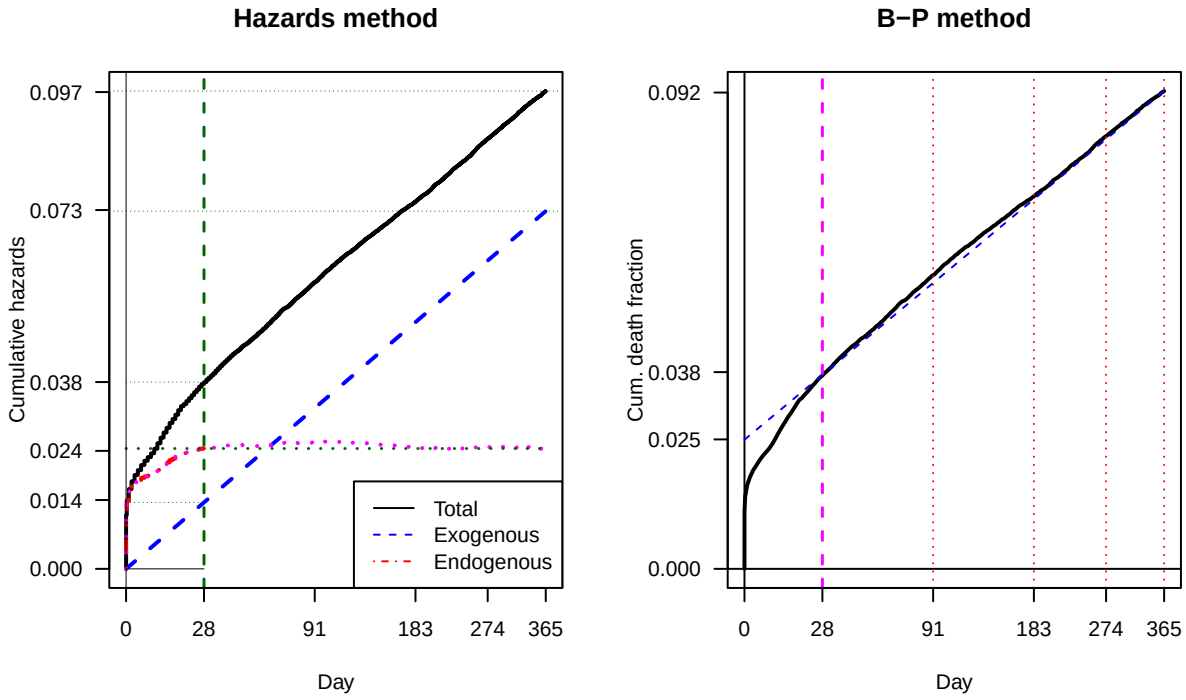

Figure 10: Västerbotten 1891–1920.

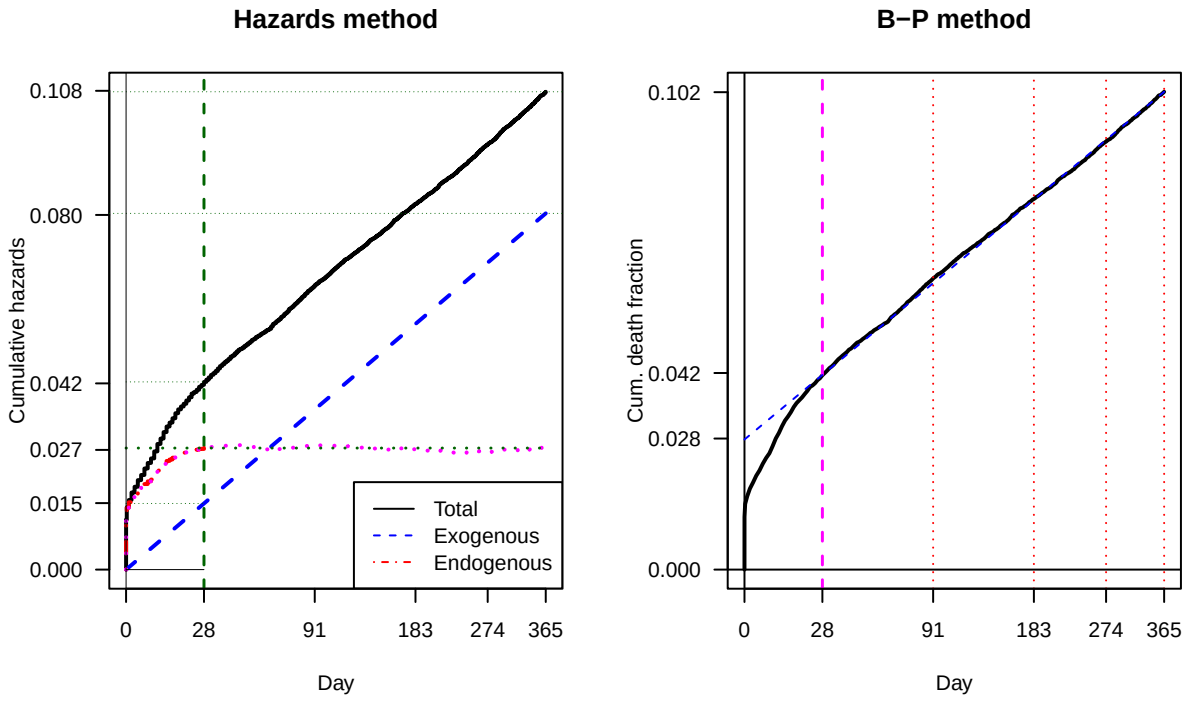

Figure 11: Västerbotten 1861–1890.

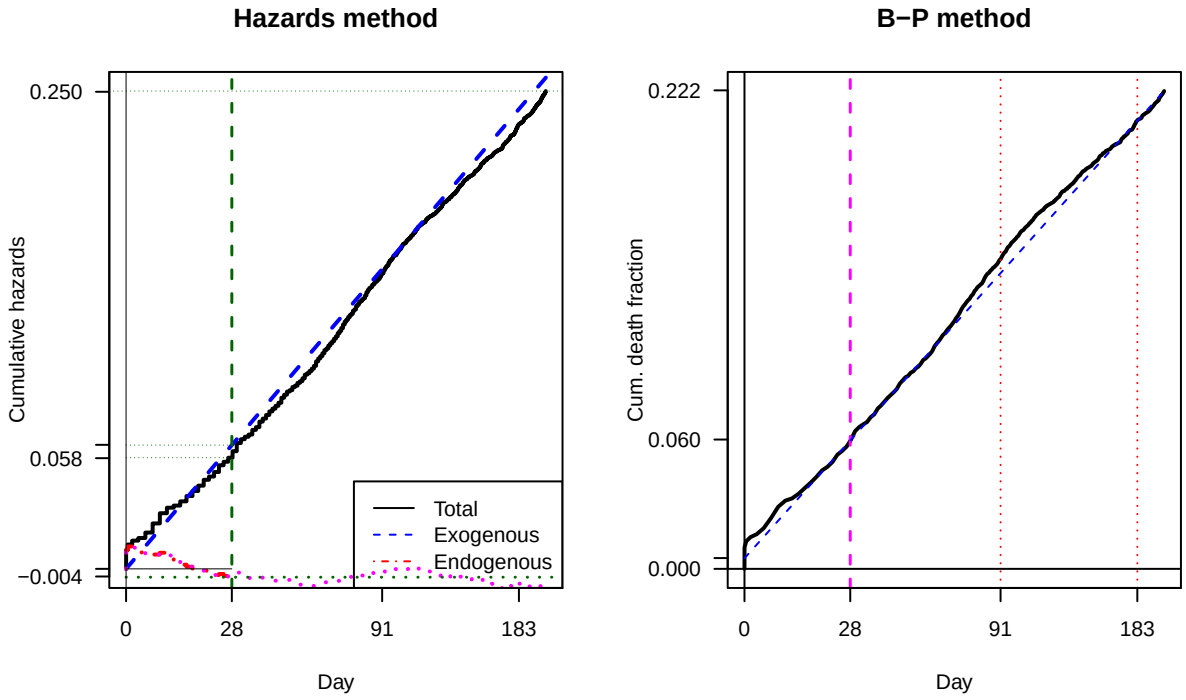

Figure 12: Scania 1710–1800.

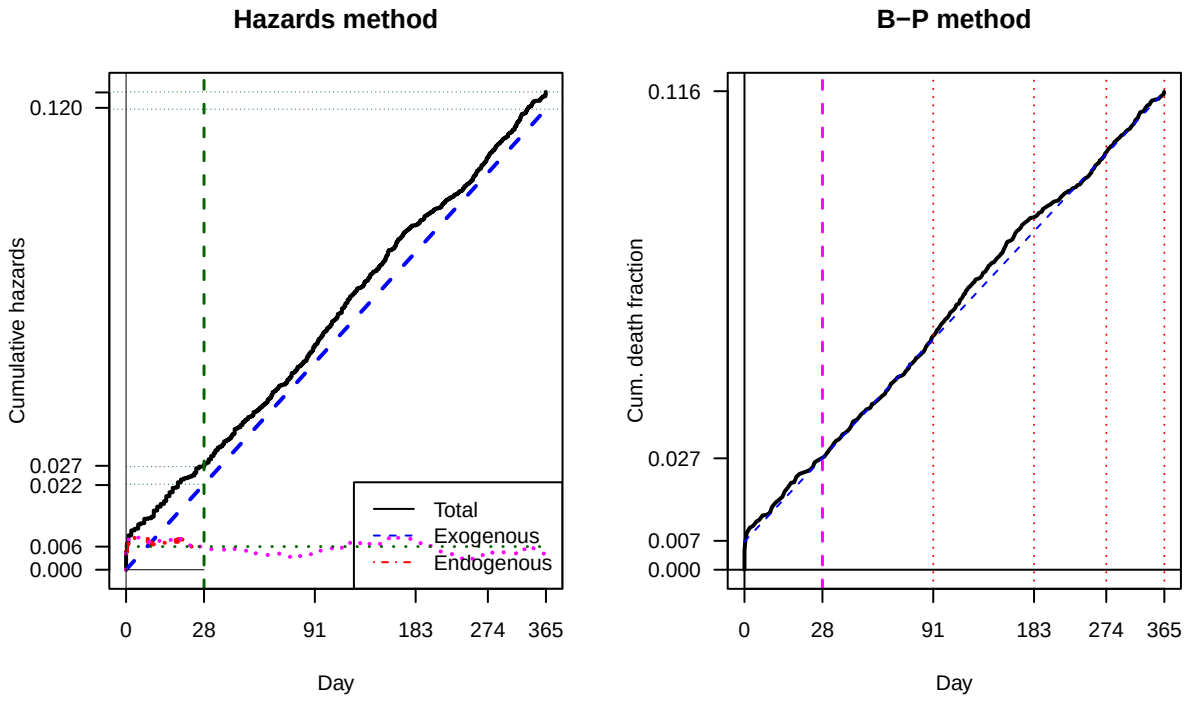

Figure 13: Skellefteå 1821–1838.