

On Determining Individual Behaviour from Population Data

MATS GYLLENBERG¹, ANDREI OSIPOV², AND LASSI PÄIVÄRINTA³

¹ Department of Mathematics, University of Turku, 20014 Turku, Finland

² Geophysical Observatory, 99600 Sodankylä, Finland

³ Department of Mathematical Sciences, University of Oulu, 90570 Oulu, Finland

Keywords: structured population dynamics, renewal equation, inverse problems.

AMS Subject Classification (2000): Primary:35R30, Secondary: 92D25.

1 Introduction

Mathematical models of *physiologically structured populations* (Metz and Diekmann 1986; Diekmann *et al.* 1998, 2001) relates mechanisms at the individual level and behaviour at the level of the population. In a typical *direct problem* one prescribes model ingredients that describe mechanisms such as aging, growth and survival at the individual level, lifts the model to the population level and finally studies phenomena at the population level. In the *inverse problem* the situation is reversed. Using knowledge about behaviour at the population level one wants to deduce the underlying mechanisms at the individual level.

The direct problem of structured populations has been extensively studied for many kinds of models using a variety of mathematical techniques (e.g. pde theory, semigroup theory, renewal theory, branching processes). We mention only the book by Webb (1985) on age-structured populations, the book by Metz and Diekmann (1986) on general physiologically structured populations and the book on branching processes by Jagers (1975). On the other hand, results on the inverse problem seem to be rare (van Straalen 1986).

In a series of papers Rundell and coworkers (Rundell 1989, 1993; Pilant and Rundell 1991b; Engl *et al.* 1994) have treated certain inverse problems of age-structured population dynamics. In these papers it is assumed that census data in the form of the age distribution of the population are available. However, for many real populations (e.g. bacterial populations) it is difficult if not impossible to measure the age distribution, whereas other quantities at the population level such as the total population size are easily obtained. Pilant and Rundell (1991a) considered in a pde setting the question of determining the initial age-distribution from data on the total population size, when the birth and death rates are known. Berndtsson and Jagers (1979) considered the same question in a branching processes framework.

In this paper we discuss the inverse problem of the simplest structured population model: the linear age-structured model (Sharpe and Lotka 1911, McKendrick 1926). We shall, in fact, use the cumulative formulation (Diekmann *et al.* 1993b, 1998; Gyllenberg *et al.* 1997) of the problem. This not only increases the generality but facilitates the analysis considerably. The specific problem we consider is: Under what conditions does knowledge of the total population size and the cumulative number of births uniquely determine the survival and reproduction functions describing individual behaviour? A typical example we have in mind is a population of cells reproducing by

fission. Here the inverse problem is indeed relevant, because the age of a cell cannot usually be measured, whereas the total number of cells is easily observed as is the fraction of dividing cells.

2 The direct problem

The classical linear age-structured population model is usually formulated as a hyperbolic pde supplemented by a nonlocal boundary condition (McKendrick 1926):

$$\frac{\partial}{\partial t}n(t, a) + \frac{\partial}{\partial a}n(t, a) = -\mu(a)n(t, a), \quad (2.1)$$

$$n(t, 0) = \int_0^\infty \beta(a)n(t, a)da, \quad (2.2)$$

$$n(0, a) = n_0(a). \quad (2.3)$$

Here the solution $n(t, \cdot)$ is the age-density of the population, $\mu(a)$ is the age-specific per capita death rate and $\beta(a)$ is the age-specific per capita fecundity.

Integrating the McKendrick equation (2.1) along characteristics one obtains

$$n(t, a) = \begin{cases} b(t-a)e^{-\int_0^a \mu(\alpha)d\alpha}, & t > a, \\ n_0(a-t)e^{-\int_{a-t}^a \mu(\alpha)d\alpha}, & t < a, \end{cases} \quad (2.4)$$

where

$$b(t) := n(t, 0)$$

is the birth rate. Substituting (2.4) into the birth law (2.2) one obtains the renewal equation

$$b(t) = \int_0^t b(t-a)\beta(a)e^{-\int_0^a \mu(\alpha)d\alpha}da + \int_t^\infty n_0(a-t)\beta(a)e^{-\int_{a-t}^a \mu(\alpha)d\alpha}da \quad (2.5)$$

for the birth rate $b(t)$. The direct problem thus reduces to solving the renewal equation (2.5) because once this has been done, (2.4) becomes an explicit formula for the solution $n(t, a)$.

Rates and densities are mathematical abstractions involving limits (derivatives) and they can never be measured. In structured population dynamics it is not even helpful to formulate models in terms of partial differential equations (Diekmann *et al.* 1993ab, 1995, 1998, 2001; Gyllenberg *et al.* 1997). We shall therefore proceed along a slightly different line, the advantages of which will be clear in a while.

From μ and β we obtain two new functions F and L by defining

$$F(a) = e^{-\int_0^a \mu(\alpha) d\alpha} \quad (2.6)$$

and

$$L(a) = \int_0^a \beta(\alpha) e^{-\int_0^\alpha \mu(\tau) d\tau} d\alpha. \quad (2.7)$$

F is called the *survival function* and L the *reproduction function*. They have the following interpretations:

$F(a)$ is the probability that an individual is still alive at age a .

$L(a)$ is the expected number of offspring born to an individual before reaching (dead or alive) age a .

Finally we introduce the cumulative number

$$B(t) := \int_{[0,t)} b(\tau) d\tau$$

of births up to time and replace the initial age density n_0 by a measure m_0 . If m_0 is absolutely continuous with respect to Lebesgue measure, then of course

$$m_0([0, a)) = \int_{[0,a)} n_0(\alpha) d\alpha$$

for some L^1 -function n_0 .

We now forget about the rates μ and β and take F and L as the basic ingredients of the population model. We take the measure m_0 as the initial population state. This increases the generality, because we can now allow discontinuities in F and L and handle initial populations described by not necessarily absolutely continuous measures. It also considerably facilitates the analysis.

The integrated versions of (2.5) and (2.4) read as

$$B(t) = \int_{[0,t)} L(t-\tau) B(d\tau) + \int_{\mathbf{R}_+} \frac{L(a+t) - L(a)}{F(a)} m_0(da) \quad (2.8)$$

and

$$m(t, [0, a)) = \int_{[0, \min\{t, a\})} F(t-\tau) B(d\tau) + \int_{[0, \max\{0, a-t\})} \frac{F(\tau+t)}{F(\tau)} m_0(d\tau), \quad (2.9)$$

respectively. The direct problem can now be formulated as follows:

(DP) Given the initial population state m_0 , the survival function F and the reproduction function L , solve the integral equation (2.8) for B . The population state $m(t, \cdot)$ at time t is then given by the explicit formula (2.9).

The interpretation of the model ingredients requires that m_0 , F and L satisfy certain conditions that we now formulate.

Assumption 2.1 The initial population state m_0 is a finite positive measure defined on \mathbf{R}_+ and the functions F and L are defined on $(0, \infty)$ and have the following properties:

- (i) F is nonnegative and nonincreasing.
- (ii) $\lim_{a \downarrow 0} F(a) = 1$.
- (iii) $F(\infty) = \lim_{a \rightarrow \infty} F(a) = 0$.
- (iv) L is nonnegative, nondecreasing, and nonlattice.
- (v) $\lim_{a \downarrow 0} L(a) = 0$.
- (vi) $R_0 = L(\infty) = \lim_{a \rightarrow \infty} L(a) < \infty$.
- (vii) $F(a) = 0 \Rightarrow L(a) = L(\infty)$,
- (viii) $\frac{1}{F} \in L^1(m_0)$.

Most of the above conditions are clear from biological point of view. However, we make the following comments: The assumption in (iv) that L is nonlattice means that L is not a step-function with discontinuities in a subset of an additive subgroup of \mathbf{R} . We thus rule out the possibility that individuals reproduce only upon exactly reaching a prescribed age a_0 (and possibly upon reaching $2a_0, 3a_0, \dots$).

If $F(A) = 0$ for some finite A , then no-one can survive beyond A and the initial measure m_0 must be concentrated on $[0, A]$. We impose the slightly stronger condition (viii), which is needed to make the latter integral in (2.9) finite. Moreover, because dead individuals do not give birth, $L(a)$ must be equal to $L(\infty)$ for ages a larger than A . This is formulated in (vii).

The number R_0 is called the *basic reproduction ratio* and it gives the expected life-time production of offspring of an individual.

The direct problem has been studied by a number of authors ever since the appearance of the paper by Sharpe and Lotka (1911). Feller (1941, 1971) finally settled all remaining open problems. Next we give a short account of those results that are important for our investigation of the inverse problem.

Assumption 2.1 ensures that the renewal equation (2.8) has a unique solution. Denote the latter term on the right hand side of (2.8) by $H(t)$. Then the Laplace transform of b is given by

$$\widehat{b}(s) = \frac{s\widehat{H}(s)}{1 - s\widehat{L}(s)} \quad (2.10)$$

Assume that the maximum life-time is finite, say 1. Then the support of F , $\text{supp } F = [0, 1]$ and $L(a) = L(1)$ for all $a \geq 1$ by Assumption 2.1. It follows that $s\widehat{H}(s)$ and $s\widehat{L}(s)$ are entire functions and hence that \widehat{b} is meromorphic in the whole plane. The zeros of $1 - s\widehat{L}(s)$ (the poles of \widehat{b}) are the roots of the *Euler-Lotka equation*

$$\int_{\mathbf{R}_+} e^{-\lambda a} L(da) = 1. \quad (2.11)$$

An easy application of Hadamard's factorization theorem (Titchmarsh 1939, p. 250) shows that the Euler-Lotka equation has infinitely many complex roots λ_k (for details, see Gyllenberg 1985). If $\widehat{b}(s)$ admits an expansion

$$\widehat{b}(s) = \sum \frac{b_k}{s - \lambda_k} \quad (2.12)$$

with $\sum |b_k| < \infty$, then, as proven by Feller (1941), the solution of (2.8) (or rather its derivative b) is representable as the series

$$b(t) = \sum b_k e^{\lambda_k t}, \quad (2.13)$$

where the series converges absolutely for all $t \geq 0$. The coefficients b_k are complex. Because b is positive the characteristic roots λ_k appear as pairs of complex conjugates. We have assumed that all roots are simple. The result above is easily generalized to the case of multiple roots (Feller 1941). Note, however, that due to positivity, the unique real root $\lambda_0 = r$ is necessarily simple and has real part larger than the real parts of all other roots. The root r is called the *Malthusian parameter* and it is positive if and only if $R_0 > 1$. For details we refer to Feller (1941, 1971).

3 The inverse problem

In many practical situations the population state, that is, the age distribution in our case, cannot be directly observed. When this is the case, it is usually also impossible to experimentally measure the model ingredients L and F . What can be observed is only certain linear functionals of the population state, called *population outputs*. The inverse problem consists of determining F and L (or μ and β) in terms of the outputs.

In this paper we shall be concerned only with two outputs, namely the *total population size*

$$N(t) = \int_{\mathbf{R}_+} m(t, da) \quad (3.1)$$

and the *population birth rate*

$$b(t) = \int_{\mathbf{R}_+} \beta(a)m(t, da). \quad (3.2)$$

Of course, a rate cannot be directly measured; the measured quantity is the *cumulative number* $B(t) = \int_{[0,t]} b(\tau)d\tau$ of births up to time t . The use of B instead of b also has the advantage that B makes sense in cases where the model is formulated in terms of the reproduction function L and not in terms of the per capita birth rate β .

It follows from (2.9) that the total population size $N(t)$ satisfies the following equation:

$$N(t) = \int_{[0,t]} F(t-\tau)B(d\tau) + \int_{\mathbf{R}_+} \frac{F(a+t)}{F(a)}m_0(da), \quad (3.3)$$

for $t > 0$. We assume that the outputs $N(t)$ and $B(t)$ are produced by admissible ingredients (that is, by functions F and L satisfying Assumption 2.1) and that they are known on a time-interval of length $T \leq \infty$. We can now give a precise formulation of the inverse problem of linear age-structured population dynamics.

(IP) Given the measure m_0 and the functions B and N defined on $[0, T)$ ($T \leq \infty$), determine the functions L and F such that the equations (3.3) and (2.8) are satisfied on $[0, T)$.

In particular, we shall be concerned with the question under which conditions the data m_0 , B and N *uniquely* determine the ingredients F and L .

On the other hand, we shall not attempt to characterize data that guarantee that F and L are admissible in the sense that they satisfy Assumption 2.1.

Once F and L have been determined the death rate μ and the fecundity β are obtained from (2.6) and (2.7) as

$$\mu(a) = -\frac{F'(a)}{F(a)}, \quad (3.4)$$

$$\beta(a) = \frac{L'(a)}{F(a)}, \quad (3.5)$$

Observe that by the monotonicity assumptions (i) and (iv) in Assumption 2.1, F and L are indeed differentiable almost everywhere, so the formulas (3.4) and (3.5) make sense.

Before we formulate our results on the inverse problem, we make a few remarks.

Remark 3.1 A naïve approach to the inverse problem would be to extend F and $b = B'$ as zero to the negative real axis and study the full line convolution equation corresponding to (3.3) with for instance Fourier transform techniques. With this convention N would be defined on the whole real axis and its support $\text{supp } N$ would be $(-\text{ess sup } m_0, \infty)$. On the other hand, the measured data contain the values of N on the positive real axis only. The failure of the naïve approach and the difficulty of the inverse problem stem from this fact.

Remark 3.2 Because the age distribution cannot in general be measured, the initial population state m_0 is in many applications unknown. It would therefore be desirable to determine not only the ingredients F and L , but also the initial age distribution m_0 from the outputs N and B . But, as shown by a simple counter example by Gyllenberg *et al.* (2002), this can never be achieved. However, our main motivation comes from the growth dynamics of cell populations and in laboratory experiments the initial population is often well synchronized and it can therefore be accurately approximated by a Dirac measure or a density with a narrow support.

Next we give a simple example, which shows that $N(t)$ and $B(t)$ need not determine the survival function $L(a)$ uniquely.

Example 3.3 Assume that $b(t)$ and $N(t)$ grow not only asymptotically, but *exactly* exponentially, that is, assume that

$$B(t) = \frac{b_0}{r} (e^{rt} - 1), \quad (3.6)$$

$$N(t) = N_0 e^{rt}, \quad (3.7)$$

and that m_0 is absolutely continuous with distributional derivative n_0 . Substituting (3.6) and (3.7) into (3.3), multiplying both sides with e^{-rt} and letting $t \rightarrow \infty$, one finds that

$$N_0 = b_0 \int_0^\infty e^{-ra} F(a) da. \quad (3.8)$$

Equation (3.3) is now an equation with F as the only unknown, and it is satisfied by

$$F(a) = \frac{1}{b_0} e^{ra} n_0(a), \quad a \geq 0. \quad (3.9)$$

With N, B and F given by (3.6), (3.7), (3.8) and (3.9) one easily checks that equation (2.8) holds for any function L satisfying

$$\int_{[0, \infty)} e^{-ra} L(da) = 1. \quad (3.10)$$

We thus conclude that the outputs (3.6) and (3.7) do *not* determine the model ingredients uniquely. Note that (3.10) is the same as the Euler–Lotka equation (2.11), but now the Malthusian parameter $\lambda = r$ is given and the survival function L has to be found.

The situation described in Example 3.3 occurs when the initial population is at demographic equilibrium. The (normalized) age-distribution remains the same for all times and it is intuitively clear that the model ingredients cannot be determined from such stable data.

At the end of Section 2 we saw that in many situations the birth rate $b(t)$ has an expansion of the form

$$b(t) = \sum b_k e^{\lambda_k t}. \quad (3.11)$$

We next show that if this expansion has only finitely many terms, then the inverse problem does not have a unique solution

Example 3.4 Assume that the birth rate $b(t)$ is given by a sum of the form (3.11), where the index k ranges over the finite set $\{0, 1, 2, \dots, q\}$ and that the total population is given by

$$N(t) = \sum_{k=0}^q N_k e^{\lambda_k t}. \quad (3.12)$$

Inserting the sums (3.11) and (3.12) into the equation (3.3) one obtains

$$\sum_{k=0}^q N_k e^{\lambda_k t} = \sum_{k=0}^q b_k e^{\lambda_k t} \int_0^\infty e^{-\lambda_k a} F(a) da, \quad (3.13)$$

which immediately implies

$$N_k = b_k \int_0^\infty e^{-\lambda_k a} F(a) da \quad \text{for all } k \in \{0, 1, 2, \dots, q\}. \quad (3.14)$$

Substituting this back into Equation (3.3) one obtains

$$\sum_{k=0}^q b_k e^{\lambda_k t} \int_0^\infty e^{-\lambda_k a} F(a) da = \sum_{k=0}^q b_k e^{\lambda_k t} \int_0^t e^{-\lambda_k a} F(a) da + \int_0^\infty \frac{n_0(a)}{F(a)} F(a+t) da \quad (3.15)$$

or

$$\int_0^\infty \left(\sum_{k=0}^q b_k e^{-\lambda_k a} - \frac{n_0(a)}{F(a)} \right) F(a+t) da = 0, \quad t \geq 0, \quad (3.16)$$

where n_0 is the distributional derivative of m_0 . Equation (3.16) is an equation in F only and it is obviously satisfied by

$$F(a) = \frac{n_0(a)}{\sum_{k=0}^q b_k e^{\lambda_k a}}. \quad (3.17)$$

As a matter of fact, (3.17) is the *unique* solution of equation (3.16), but this is irrelevant for this example.

Substituting (3.17) and (3.11) into (2.8) one finds that every function L satisfying

$$\int_{[0, \infty)} e^{-\lambda_k a} L(da) = 1 \quad \text{for all } k \in \{0, 1, 2, \dots, q\} \quad (3.18)$$

is a solution of equation (2.8). We shall show that if there exists one such function L satisfying Assumption 2.1, then there are in fact infinitely many such functions L_ε satisfying Assumption 2.1.

View L as a positive measure on $[0, 1]$ and let K be the support of L . Because the functions $e^{-\lambda_k a}$ belong to $L^1(L)$, the Hahn-Banach theorem guarantees the existence of a nonzero measurable and essentially bounded (with respect to the measure L) function g on K such that

$$\int_K e^{-\lambda_k a} g(a) L(da) = 0 \quad \text{for all } k \in \{0, 1, 2, \dots, q\}. \quad (3.19)$$

Extend g to all of $[0, 1]$ by defining

$$g(a) = 0 \quad \text{for } a \in [0, 1] \setminus K. \quad (3.20)$$

Then the measure

$$L_\varepsilon(da) = (\varepsilon g(a) + 1)L(da) \quad (3.21)$$

is a positive measure on $[0, 1]$ for all $0 < \varepsilon < 1/\|g\|_\infty$ and it satisfies (3.18) because g satisfies (3.19) and L satisfies (3.18).

The situation changes drastically when the sum in (3.11) involves infinitely many terms. Our main theorem (Theorem 3.5) says that the inverse problem has a unique solution whenever the exponential functions $e^{\lambda_k t}$ are complete in the continuous functions.

Theorem 3.5 *Let $\text{supp } m_0 \subset [0, 1]$ and assume that m_0 does not have an atom at 1. Let $b(t)$ and $N(t)$ be defined on $[0, 1 + \varepsilon)$ and representable as series*

$$b(t) = \sum b_k e^{\lambda_k t} \quad (3.22)$$

and

$$N(t) = \sum N_k e^{\mu_k t} \quad (3.23)$$

which converge absolutely for $t \in [0, 1 + \varepsilon)$, where $\varepsilon > 0$. Then Equation (3.3) has a solution F with support $[0, 1]$ if and only if

$$\lambda_k = \mu_k \quad \text{for all } k, \quad (3.24)$$

and

$$N_k = b_k \int_0^1 e^{-\lambda_k a} F(a) da \quad \text{for all } k. \quad (3.25)$$

If this is the case, then the function $b(t)$ can be real-analytically continued to $(-1, 1 + \varepsilon)$. Moreover, the solution F of (3.3) is unique and given by

$$F(a) = \frac{n_0(a)}{b(-a)}, \quad (3.26)$$

where n_0 is the distributional derivative of m_0 . Moreover, every function L for which

$$\int_{[0,1]} e^{-\lambda_k a} L(da) = 1 \quad \text{for all } k \quad (3.27)$$

satisfies equation (2.8). There is a unique such L if and only if

$$\sum \left(1 - \left| \frac{\lambda_k + 2\pi i}{\lambda_k - 2\pi i} \right| \right) = \infty. \quad (3.28)$$

For a complete proof of this theorem we refer to Gyllenberg *et al.* (2002)

Acknowledgement The research of Mats Gyllenberg and Lassi Päivärinta has been supported by the Academy of Finland.

References

- BERNDTSSON, B. and JAGERS, P. (1979) Exponential growth of a branching process usually implies stable age distribution. *J. Appl. Prob.* **16** 651-656.
- DIEKMANN, O., GYLLENBERG, M., and THIEME, H.R. (1993a) Perturbing semigroups by solving Stieltjes renewal equations. *Differential and Integral Equations* **6** 155-181.
- DIEKMANN, O., GYLLENBERG, M., METZ, J.A.J., and THIEME, H.R. (1993b) The “cumulative” formulation of (physiologically) structured population models, In “Evolution Equations, Control Theory and Biomathematics”, (Ph. Clément and G. Lumer, Eds.), pp. 145-154, Marcel Dekker, New York.
- DIEKMANN, O., GYLLENBERG, M., and THIEME, H.R. (1995) Perturbing evolutionary systems by step responses and cumulative outputs. *Differential and Integral Equations* **8** 1205-1244.
- DIEKMANN, O., GYLLENBERG, M., METZ, J.A.J., and THIEME, H.R. (1998) On the formulation and analysis of general deterministic structured population models. I. Linear theory. *Journal of Mathematical Biology* **36** 349-388.
- DIEKMANN, O., GYLLENBERG, M., HUANG, H., KIRKILIONIS, M., METZ, J.A.J., and THIEME, H.R. (2001) On the formulation and analysis of general deterministic structured population models. II. Nonlinear theory. *Journal of Mathematical Biology* **43** 157-189.
- ENGL, H.W., RUNDELL, W., SCHERZER, O. (1994) A regularization scheme for an inverse problem in age-structured populations. *J. Math. Anal. Appl.* **182** 658-679.
- FELLER, W. (1941) On the integral equation of renewal theory, *Ann. Math. Statist.* **12** 243-267.
- FELLER, W. (1971) *An introduction to probability theory and its applications* Vol.II, Second Edition, Wiley, New York.
- GYLLENBERG, M. (1985) The age structure of populations of cells reproducing by asymmetric division, In *Mathematics in Biology and Medicine*, V. Capasso, E. Grosso, and S.L. Paveri-Fontana (Eds.), Springer, Berlin, pp. 320-327.
- GYLLENBERG, M., HANSKI, I., and HASTINGS, A. (1997) Structured metapopulation models. In: “Metapopulation biology: ecology, genetics and evo-

lution” (I.A. Hanski and M.E. Gilpin, Eds.) pp. 93–122, Academic Press, San Diego.

GYLLENBERG, M., OSIPOV, A. and PÄIVÄRINTA, L. (2002) An inverse problem of age-structured population dynamics, *Journal of Evolution Equations*, in press.

JAGERS, P. (1975) *Branching Processes with Biological Applications*, Wiley, London.

MCKENDRICK, A.G. (1926) Applications of Mathematics to Medical Problems, *Proc. Edinb. Math. Soc.* **44** 98–130.

METZ, J.A.J. and DIEKMANN, O. (1986) *The Dynamics of Physiologically Structured Populations*, Springer, Berlin.

PILANT, M. AND RUNDELL, W. (1991a) Determining the initial age distribution for an age structured population, *Math. Population Stud.* **3** 3-20.

PILANT, M. AND RUNDELL, W. (1991b) Determining a coefficient in a first-order hyperbolic equation, *SIAM J. Appl. Math.* **51** 294–506.

RUNDELL, W. (1989) Determining the birth function for an age structured population. *Math. Population Stud.* **1** 377–395, 397.

RUNDELL, W. (1993) Determining the death rate for an age-structured population from census data. *SIAM J. Appl. Math.* **53** 1731–1746.

SHARPE, F.R. and LOTKA, A.J. (1911) A problem in age-distribution, *Philosophical Magazine* **21** 435–438.

VAN STRAALLEN, N.M. (1986) The “inverse problem” in demographic analysis of stage-structured populations, In *The Dynamics of Physiologically Structured Populations*, J.A.J Metz and O. Diekmann (Eds.) pp. 393–408, Springer, Berlin.

TITCHMARSH, E.C. (1939) *The Theory of Functions*. Second Edition. Oxford University Press, Glasgow.

WEBB, G.F. (1985) *Nonlinear Age-Dependent Population Dynamics*, Marcel Dekker, New York.