Appendix A

A.1 Derivation of formulas in section 3.6.2 for Sykes (1969)

The formula for $\mathrm{var}(\mathrm{P}(t))$ for the additive errors model can be derived in the following way

$$\begin{aligned} Cov(\mathrm{P}(t),\mathrm{P}(u)) &= E\left[(\mathrm{P}(t)-E(\mathrm{P}(t)))(\mathrm{P}(u)-E(\mathrm{P}(u))^{'}\right] \\ &= \left[\left(\sum_{i=1}^{t}\mathrm{A}^{i}S_{t-i-1}\right)\left(\sum_{j=1}^{u}\mathrm{A}^{j}S_{u-j-1}\right)\right] \\ &= E\left[\sum_{i=1}^{t}\sum_{j=1}^{u}\mathrm{A}^{i}S_{t-i-1}S^{'}{}_{u-j-1}\mathrm{A}^{'j}\right] \\ &= \sum_{i=1}^{t}\sum_{j=1}^{u}\mathrm{A}^{i}\Psi_{t-i-1,u-j-1}\mathrm{A}^{'j} \end{aligned}$$

That indicates that $\mathrm{var}(\mathrm{P}(t))$ is given by

$$Var(\mathrm{P}(t)) = \sum_{i=1}^{t}\sum_{j=1}^{t}\mathrm{A}^{i}\Psi_{t-i-1,t-j-1}\mathrm{A}^{'j}$$

For the Branching process formulation Sykes suggests using a conditional argument. The expectation can be given by

$$\begin{aligned} E(\mathrm{P}(t+1)) &= E\{E(\mathrm{P}(t+1)\mid\mathrm{P}(t))\} \\ &= E(\mathrm{AP}(t)) \\ &= \mathrm{A}E(\mathrm{P}(t)) \\ &= \mathrm{A}^{t}\mathrm{P}(0) \end{aligned}$$

The conditional variance formula is given by

$$Var(\mathrm{P}(t+1)) = E(Var\mathrm{P}(t+1)\mid\mathrm{P}(t)) + Var(E(\mathrm{P}(t+1)\mid\mathrm{P}(t)))$$

Applying the conditional variance formula leads to

$$\begin{aligned} (Var\mathrm{P}(t+1)\mid\mathrm{P}(t)) &= \mathrm{var}(\mathrm{AP}(t))\mid\mathrm{P}(t) \\ &= \mathrm{A}\,\mathrm{var}(\mathrm{P}(t)) \\ &= E\left([\mathrm{AP}(t)][\mathrm{AP}(t)]^{'}\right) - (E[\mathrm{AP}(t)])^{2} \\ &= E\left(\sum_{\alpha}\sum_{\beta}a_{i\alpha}a_{i\beta}\mathrm{P}_{\alpha}(t)\mathrm{P}_{\beta}(t)\right) - (E(\mathrm{AP}(t)))^{2} \\ &= \left(\sum_{\alpha}\sum_{\beta}a_{i\alpha}a_{i\beta}E(\mathrm{P}_{\alpha}(t)\mathrm{P}_{\beta}(t))\right) - (\mathrm{A}E(\mathrm{P}(t)))^{2} \end{aligned}$$

$$
\begin{aligned}
&= \left(\sum_{\alpha}\sum_{\beta} a_{i\alpha}a_{i\beta}\left[\operatorname{cov}\left(\mathrm{P}_{\alpha}(t),\mathrm{P}_{\beta}(t)\right)+ E\left(\mathrm{P}_{\alpha}(t)\right)E\left(\mathrm{P}_{\beta}(t)\right)\right]\right) - \left(\mathrm{A}E\left(\mathrm{P}(t)\right)\right)^2 \\
&= \sum_{\alpha}\sum_{\beta} a_{i\alpha}a_{i\beta}\left[\operatorname{cov}\left(\mathrm{P}_{\alpha}(t),\mathrm{P}_{\beta}(t)\right)\right]+ \sum_{\alpha}\sum_{\beta} a_{i\alpha}a_{i\beta}E\left(\mathrm{P}_{\alpha}(t)\right)E\left(\mathrm{P}_{\beta}(t)\right) - \\
&\quad \left(\mathrm{A}E\left(\mathrm{P}(t)\right)\right)^2 \\
&= \sum_{\alpha}\sum_{\beta} a_{i\alpha}a_{i\beta}\left[\operatorname{cov}\left(\mathrm{P}_{\alpha}(t),\mathrm{P}_{\beta}(t)\right)\right]
\end{aligned}
$$

It follows that

$$
\operatorname{var}\left(\mathrm{P}(t)\right) = \sum_{\alpha}\sum_{\beta} a_{i\alpha}a_{i\beta}\left[\operatorname{cov}\left(\mathrm{P}_{\alpha}(t),\mathrm{P}_{\beta}(t)\right)\right]+ \mathrm{A}\operatorname{var}\left(\mathrm{P}(t-1)\right)\mathrm{A}'
$$

This can be written as

$$
\operatorname{var}\left(\mathrm{P}(t)\right)= \sum_{i=0}^{t-1} \mathrm{A}^{t-1-i} C_i \mathrm{A}'^{t-1-i},
$$

where

$$
C_i = \sum_{\alpha}\sum_{\beta} a_{i\alpha}a_{i\beta}\left[\operatorname{cov}\left(\mathrm{P}_{\alpha}(i),\mathrm{P}_{\beta}(i)\right)\right].
$$

Sykes' third approach is that of the random transition matrix. The author introduces the model

$$
\mathrm{P}(t+1) = \left(\mathrm{A} + \mathrm{K}(t)\right)\mathrm{P}(t),\ t = 0,1,\ldots,
$$

where $\mathrm{K}(t)$ is a sequence of independent $l \times l$ matrix random variables satisfying $E\left(\mathrm{K}(t)\right) = 0$, and $\operatorname{var}\left(\mathrm{K}(t)\right) = \Sigma$, with Σ is a singular $l^2 \times l^2$ matrix. To find $E\left(\mathrm{P}(t+1)\right)$ and $\operatorname{var}\left(\mathrm{P}(t+1)\right)$ Sykes uses the conditional mean to obtain the expectation as follows

$$
\begin{aligned}
E\left(\mathrm{P}(t+1) \mid \mathrm{P}(t)\right) &= E\left(\mathrm{A} + \mathrm{K}(t)\right)\mathrm{P}(t) \mid \mathrm{P}(t) \\
&= \mathrm{A}\mathrm{P}(t) + E\left(\mathrm{K}(t)\mathrm{P}(t)\right) \mid \mathrm{P}(t) \\
&= \mathrm{A}\mathrm{P}(t) + \mathrm{P}(t)E\left(\mathrm{K}(t) \mid \mathrm{P}(t)\right) \\
&= \mathrm{A}\mathrm{P}(t) + \mathrm{P}(t)E\left(\mathrm{K}(t)\right), \text{ because of independence} \\
&= \mathrm{A}\mathrm{P}(t) \\
\Rightarrow \quad E\left(\mathrm{P}(t+1)\right) &= E\left\{E\left(\mathrm{P}(t+1) \mid \mathrm{P}(t)\right)\right\} \\
&= E\left(\mathrm{A}\mathrm{P}(t)\right) \\
&= \mathrm{A}E\left(\mathrm{P}(t)\right) \\
&= \mathrm{A}^t\mathrm{P}(0)
\end{aligned}
$$

Sykes makes use of the conditional variance formula to calculate the variance as follows

$$
\begin{aligned}
Var(\mathrm{P}(t+1)) &= Var\{(\mathrm{A}+\mathrm{K}(t))\mathrm{P}(t) \mid \mathrm{P}(t)\} \\
&= Var\{\mathrm{K}(t)\mathrm{P}(t) \mid \mathrm{P}(t)\} \\
&= E(\mathrm{K}(t)\mathrm{P}(t))(\mathrm{K}(t)\mathrm{P}(t))' \mid \mathrm{P}(t) - \{E(\mathrm{K}(t))E(\mathrm{P}(t))\}^2 \\
&= E\{\mathrm{K}(t)\mathrm{P}(t)\mathrm{P}'(t)\mathrm{K}'(t)\} \mid \mathrm{P}(t) \\
&= E\left[\left(\sum_{v=1}^{p}\sum_{w=1}^{p}\mathrm{K}_v(t)\mathrm{K}_w(t)\mathrm{P}_v(t)\mathrm{P}_w(t)\right) \mid \mathrm{P}(t)\right]
\end{aligned}
$$

where the curly brackets indicate that expression inside them is the ij^{th} element of the matrix considered. It follows that

$$
Var(\mathrm{P}(t+1) \mid \mathrm{P}(t)) = \sum_{v=1}^{l}\sum_{w=1}^{l}\mathrm{cov}(\mathrm{K}_v(t)\mathrm{K}_w(t))\mathrm{P}_v(t)\mathrm{P}_w(t)
$$

This implies that

$$
\begin{aligned}
Var(\mathrm{P}(t+1)) &= \sum_{v=1}^{l}\sum_{w=1}^{l}\mathrm{cov}(\mathrm{K}_v(t)\mathrm{K}_w(t))E(\mathrm{P}_v(t)\mathrm{P}_w(t)) + Var(\mathrm{AP}(t)) \\
&= \sum_{v=1}^{l}\sum_{w=1}^{l}\mathrm{cov}(\mathrm{K}_v(t)\mathrm{K}_w(t))[\mathrm{cov}(\mathrm{P}_v(t),\mathrm{P}_w(t)) + E(\mathrm{P}_v(t))E(\mathrm{P}_w(t))] + \\
&\quad Var(\mathrm{AP}(t)) \\
&= \sum_{v=1}^{l}\sum_{w=1}^{l}\mathrm{cov}(\mathrm{K}_v(t)\mathrm{K}_w(t))[\mathrm{cov}(\mathrm{P}_v(t),\mathrm{P}_w(t)) + E(\mathrm{P}_v(t))E(\mathrm{P}_w(t))] + \\
&\quad \mathrm{A}Var(\mathrm{P}(t))\mathrm{A}'
\end{aligned}
$$

That indicates that

$$
\begin{aligned}
Var\mathrm{P}(t) = \sum_{v=1}^{l}\sum_{w=1}^{l}\mathrm{cov}(\mathrm{K}_v(t-1)\mathrm{K}_w(t-1))[&\mathrm{cov}(\mathrm{P}_v(t-1),\mathrm{P}_w(t-1)) + \\
&E(\mathrm{P}_v(t-1))E(\mathrm{P}_w(t-1)) + \mathrm{A}Var(\mathrm{P}(t-1))\mathrm{A}']
\end{aligned}
$$

Let

$$
\mathrm{O}(t) = \sum_{v=1}^{l}\sum_{w=1}^{l}\mathrm{cov}(\mathrm{K}_v(t)\mathrm{K}_w(t))[\mathrm{cov}(\mathrm{P}_v(t),\mathrm{P}_w(t)) + E(\mathrm{P}_v(t))E(\mathrm{P}_w(t))]
$$

Then

$$
\begin{aligned}
Var\mathrm{P}(t) &= \mathrm{O}(t-1) + \mathrm{A}Var(\mathrm{P}(t-1))\mathrm{A}' \\
&= \sum_{i=0}^{t-1}\mathrm{A}^{t-1-i}\mathrm{O}_i\mathrm{A}'^{t-1-i}
\end{aligned}
$$

Writing out O gives

$$Var\mathrm{P}(t) = \sum_{l=0}^{t-1} \mathrm{A}^{t-1-l} \sum_{v=1}^{l}\sum_{w=1}^{l} \mathrm{cov}\left(\mathrm{K}_v(t)\mathrm{K}_w(t)\right)\left[\mathrm{cov}\left(\mathrm{P}_v(t),\mathrm{P}_w(t)\right) + E\left(\mathrm{P}_v(t)\right)E\left(\mathrm{P}_w(t)\right)\right]\mathrm{A}^{t-1-l}$$

A.2 Derivation of the formula of $d(t)$ in section 3.6.2 for Lee (1974)

Equation 3.2.1 can be written in the following way

$$\begin{aligned} B(t) &= B + Bd(t) \\ &= \sum_{j=1}^{Upper} \left(n(j) - x(j,t)\right)\left(B + Bd(t-j)\right) \\ &= B\sum_{j=1}^{Upper} \left(n(j) + n(j)d(t-j) + x(j,t) + x(j,t)d(t-j)\right) \end{aligned}$$

That means that

$$\begin{aligned} d(t) &= \sum_{j=1}^{Upper} n(j) + \sum_{j=1}^{Upper} n(j)d(t-j) + x(j,t) + x(j,t)d(t-j) - 1 \\ &= \sum_{j=1}^{Upper} n(j)d(t-j) + \sum_{j=1}^{Upper} \left(n(j,t) - n(j)\right) + \sum_{j=1}^{Upper} x(j,t)d(t-j) \\ &= \sum_{j=1}^{Upper} n(j)d(t-j) + \varepsilon(t) + \sum_{j=1}^{Upper} x(j,t)d(t-j) \end{aligned}$$

Based on earlier results Lee ignores the last term and approximates $d(t)$ by the following AR process

$$d(t) = \sum_{j=1}^{Upper} n(j)d(t-j) + \varepsilon(t)$$

A.3 Derivation of formulas in section 3.6.2 for Alho and Spencer (1991)

The covariance between the prediction error of population size in ages i and j at time t, $1 \le t \le i \le j \le l$, is given by

$$\mathrm{cov}\left(\tilde{p}(i,t), \tilde{p}(j,t)\right) = \mathrm{cov}\left[\left(\tilde{p}(i-t,0) + \sum_{m=0}^{t-1} \tilde{sv}(i-t+m,m)\right), \left(\tilde{p}(j-t,0) + \sum_{n=0}^{t-1} \tilde{sv}(j-t+n,n)\right)\right]$$

$$= \operatorname{cov}\left[\tilde{p}(i-t,0), \tilde{p}(j-t,0)\right] + \operatorname{cov}\left[\sum_{m=0}^{t-1}\tilde{sv}(i-t+m,m), \sum_{n=0}^{t-1}\tilde{sv}(j-t+n,n)\right]$$

$$= \sigma(i,j,t) + \sum_{n=0}^{t-1}\sum_{m=0}^{t-1}\sigma_{sv}(i-t+m, j-t+n, m, n)$$

The Taylor series expansion used by Alho and Spencer (1991) can be described as follows. Consider the function

$$h(y_1,...,y_n) = \log\left(\sum_{i=1}^{n}\exp(y_i)\right)$$

Then a linear Taylor series representation for h at $z_1,...,z_n$ is given by

$$h_L(y_1,...,y_n) = h(z_1,...,z_n) + \exp\left(-h(z_1,...,z_n)\right) * \sum_{i=1}^{n}\exp(z_i)(y_i - z_i)$$

Alho and Spencer apply Taylor series expansion by taking $y_i = \log\left(\tilde{\mathrm{B}}(j,t)\right)$ and $z_i = \log\left(\mathrm{B}(j,t)\right)$. Therefore

$$h(z_1,...,z_n) = \log\sum_{j=15}^{44}\exp\{\log \mathrm{B}(j,t)\} = \log\sum_{j=15}^{44}\mathrm{B}(j,t) = p(0,t),$$

$\exp\left(-h(z_1,...,z_n)\right) = 1/\,p(0,t)$, and

$$\sum_{j=15}^{44}\exp(z_j)(y_j - z_j) = \sum_{j=15}^{44}\mathrm{B}(j,t)\left(\log(\tilde{\mathrm{B}}(j,t)) - \log(\mathrm{B}(j,t))\right)$$

$$\Rightarrow \tilde{p}(0,t) = p(0,t) + \frac{1}{\mathrm{P}(0,t)}\left\{\sum_{j=15}^{44}\mathrm{B}(j,t)\left(\log(\tilde{\mathrm{B}}(j,t)) - \log(\mathrm{B}(j,t))\right)\right\}$$

The covariance between the jump-off populations at two different times is given by

$$\sigma(0,0,t,u) \approx \operatorname{cov}\left(\left(1/\mathrm{P}(0,t)\right)\sum_{i=15}^{44}\mathrm{B}(i,t)\left(\log(\tilde{\mathrm{B}}(i,t)) - \log(\mathrm{B}(i,t))\right), \left(1/\mathrm{P}(0,u)\right)\sum_{j=15}^{44}\mathrm{B}(j,u)\right.$$

$$\left.\left(\log(\tilde{\mathrm{B}}(j,u)) - \log(\mathrm{B}(j,u))\right)\right)$$

$$\approx \left(1/\mathrm{P}(0,t)\,1/\mathrm{P}(0,u)\right)\sum_{i=15}^{44}\sum_{j=15}^{44}\mathrm{B}(i,t)\mathrm{B}(j,u)\left[\operatorname{cov}\left(\log\tilde{\mathrm{B}}(i,t)\right), \left(\log\tilde{\mathrm{B}}(j,u)\right)\right]$$

Since

$$\tilde{\mathrm{B}}(k,t) = \exp\left(\tilde{p}(k-t+1,0) + \sum_{n=0}^{t-2} \tilde{sv}(k-t+1+n,n) + \tilde{ft}(k,t-1) \right)$$

it follows that, using the independence assumption between the jump-off population and vital rates and between fertility and survival rates,

$$\sigma(0,0,t,u) \approx \left(1 \Big/ \mathrm{P}(0,t)\mathrm{P}(0,u) \right) \sum_{i=15}^{44} \sum_{j=15}^{44} \mathrm{B}(j,t)\mathrm{B}(j,t)(\mathrm{cov}(\tilde{p}(i-t+1,0), \tilde{p}(j-u+1,0)) +$$

$$\mathrm{cov} \sum_{m=0}^{t-2} \tilde{sv}(i-t+1+m,m), \sum_{n=0}^{t-2} \tilde{sv}(j-u+1+n,n) + \mathrm{cov}(\tilde{ft}(i,t-1), \tilde{ft}(j,u-1))$$

Therefore, $\sigma(0,0,t,u)$ is given by

$$\sigma(0,0,t,u) \approx \left(1 \Big/ \mathrm{P}(0,t)\mathrm{P}(0,u) \right) \sum_{i=15}^{44} \sum_{j=15}^{44} \mathrm{B}(j,t)\mathrm{B}(j,t)(\sigma(i-t+1, j-u+1) +$$

$$\sum_{m=0}^{t-2} \sum_{n=0}^{u-2} \sigma_{sv}(i-t+1+m, j-u+1+n, m, n) + \sigma_{ft}(i, j, t-1, u-1))$$

For the surviving births, $\max[0, t-16] \le j \le i < t$, $\mathrm{cov}\left(\tilde{p}(i,t), \tilde{p}(j,t) \right)$, the covariance is given by

$$\mathrm{cov}\left(\tilde{p}(i,t), \tilde{p}(j,t) \right) = \mathrm{cov}\left\{ \tilde{p}(0,t-i) + \sum_{n=0}^{i-1} \tilde{sv}(n,t-i+n), \tilde{p}(0,t-j) + \sum_{m=0}^{j-1} \tilde{sv}(m,t-j+m), \right\}$$

$$= \sigma(0,0,t-i,t-j) + \sum_{n=0}^{i-1} \sum_{m=0}^{j-1} \sigma_{sv}(n,m,t-i+n,t-j+m) +$$

$$\mathrm{cov}\left(\tilde{p}(0,t-i), \sum_{m=0}^{j-1} \tilde{sv}(m,t-j+m \right) + \mathrm{cov}\left(\tilde{p}(0,t-j), \sum_{n=0}^{j-1} \tilde{sv}(n,t-i+n \right)$$

Using the Taylor series expansion explained earlier

$$\mathrm{cov}\left(\tilde{p}(0,t-i), \sum_{m=0}^{j-1} \tilde{sv}(m,t-j+m) \right) = \mathrm{cov}\left(\left(1 \Big/ \mathrm{P}(0,t-i) \right) \sum_{k=15}^{44} \mathrm{B}(k,t-i) \left(\log\left(\tilde{\mathrm{B}}(k,t-i) \right) - \right. \right.$$

$$\log\left(\mathrm{B}(k,t-i)\right), \sum_{m=0}^{j-1} \tilde{sv}(m,t-j+m)\Big)$$
$$\approx \left(1/\mathrm{P}(0,t-i)\right) \sum_{k=15}^{44} \mathrm{B}(k,t-i)\,\mathrm{cov}\left(\log\left(\tilde{\mathrm{B}}(k,t-i)\right), \sum_{m=0}^{j-1} \tilde{sv}(m,t-j+m)\right)$$

But

$$\tilde{\mathrm{B}}(k,t-i) = \exp\left(\tilde{p}(k-t+i+1,0) + \sum_{n=0}^{t-i-2} \tilde{sv}(k-t+i+1+n,n) + \tilde{ft}(k,t-i-1)\right)$$

Therefore

$$\mathrm{cov}\left(\log\left(\tilde{\mathrm{B}}(k,t-i)\right), \sum_{m=0}^{j-1} \tilde{sv}(m,t-j+m)\right) \approx \mathrm{cov}\left(\tilde{p}(k-t+i+1,0) + \sum_{n=0}^{t-i-2} \tilde{sv}(k-t+i+1+n,n) + \tilde{ft}(k,t-i-1), \sum_{m=0}^{j-1} \tilde{sv}(m,t-j+m)\right)$$
$$\approx 1/\mathrm{P}(0,t-i) \sum_{k=15}^{44} \mathrm{B}(k,t-i) \sum_{n=0}^{t-i-2} \sum_{m=0}^{j-1} \sigma_{sv}[k-t+i+1+n,m,n,t-j+m]$$

That means that

$$\mathrm{cov}\left(\tilde{p}(i), \tilde{p}(j)\right) \approx \sigma(0,0,t-i,t-j) + \sum_{m}^{i-1} \sum_{n}^{j-1} \sigma_{sv}(m,n,t-i+m,t-j+n) + 1/\mathrm{P}(0,t-i) \sum_{k=15}^{44} \mathrm{B}(k,t-i) \sum_{n=0}^{t-i-2} \sum_{m=0}^{j-1} \sigma_{sv}[k-t+i+1+n,m,n,t-j+m]$$

(A.3.1)

Alho and Spencer set the covariance between the survival rates of the mothers giving birth at time t and the survival rates of their own mothers (17 or more years earlier) to zero, leading to the cancellation of the last term in the formula above.

The covariance between the surviving births in age i, $\max\{0,t-16\} \le i < t$, at time t and the survivors of the jump-off population at age j at time u $(u \le j)$ is given by

$$\mathrm{cov}\left(\tilde{p}(i,t), \tilde{p}(j,u)\right) = \mathrm{cov}[\tilde{p}(0,t-i) + \sum_{n=0}^{i-1} \tilde{sv}(n,t-i+n), \tilde{p}(j-u,0) + \sum_{m=0}^{u-1} \tilde{sv}(j-u+m,m)$$
$$= \mathrm{cov}\left[\tilde{p}(0,t-i), \tilde{p}(j-u,0)\right] +$$
$$\mathrm{cov}\left[\sum_{n=0}^{i-1} \tilde{sv}(n,t-i+n), \sum_{m=0}^{u-1} \tilde{sv}(j-u+m,m)\right] +$$

$$\operatorname{cov}\left[\tilde{p}(0,t-i),\sum_{m=0}^{u-1}\tilde{sv}(j-u+m,m)\right]+\operatorname{cov}\left[\sum_{n=0}^{i-1}\tilde{sv}(n,t-i+n),\tilde{p}(j-u,0)\right]$$

Using Taylor series expansion, $\tilde{p}(0,t-i)$ can be written as

$$\tilde{p}(0,t-i)\approx p(0,t-i)+\left(1/\mathrm{P}(0,t-i)\right)\sum_{k=15}^{44}\mathrm{B}(k,t-i)\left[\log(\tilde{\mathrm{B}}(k,t-i))-\log(\mathrm{B}(k,t-i))\right]$$

with

$$\tilde{\mathrm{B}}(k,t-i)=\exp\left(\tilde{p}(k-t+i+1,0)+\sum_{r=0}^{t-i-2}\tilde{sv}(k-t+i+1+r,r)+\tilde{ft}(k,t-i-1)\right)$$

Therefore, the covariance between the surviving births in age i at time t and the survivors of the jump-off population at age j at time $u\,(u\le j)$ can be given by

$$\operatorname{cov}\left(\tilde{p}(i,t),\tilde{p}(j,u)\right)\approx\sum_{n=0}^{i-1}\sum_{m=0}^{n-1}\sigma_{sv}(n,j-u+m,t-i+n,m)+$$

$$\left(1/\mathrm{P}(0,t)\right)\left(\sum_{k=15}^{44}\mathrm{B}(k,t-i)[\sigma(k-t+i+1,j-u,0)+\right.$$

$$\left.\sum_{r=0}^{t-i-2}\sum_{m=0}^{u-1}\sigma_{sv}(k-t+i+1+r,j-u+m,r,m)]\right)\quad\text{(A.3.2)}$$

Alho and Spencer set the covariance between the errors in the jump-off population and the errors in births of the mothers giving birth at time t to zero. The authors also set the covariance between the errors in the fertility forecasts for year $t-1$ and the past births to zero. These simplifications lead to the cancellation of the last term in A.3.2.

A.4 Derivation of formulas in section 3.6.2 for Alho (1992a) and Alho (1992b)

Given $V(t)=\sum_{k=0}^{t-1}\left(\varepsilon_{sv}(k)+\varepsilon_{mg}(k)\right)$, Alho starts with evaluating the term $\sum_{k=1}^{t-1}\varepsilon_{sv}(k)$ by noting that

$$\sum_{k=1}^{t-1}\varepsilon_{sv}(k)=\sum_{k=0}^{t-1}(t-k)e_{sv}(k)$$

$$\Rightarrow\operatorname{var}\left(\sum_{k=0}^{t-1}\varepsilon_{sv}(k)\right)=\sigma_{sv}{}^{2}\sum_{k=0}^{t-1}(t-k)^{2}=\sigma_{sv}{}^{2}g(t)$$

where $g(t) = (2t+1)(t+1)t/6$. It follows that

$$\text{var}(\varepsilon_{JO} + V(t)) = {\sigma_{JO}}^2 + {\sigma_{sv}}^2 g(t) + t{\sigma_{mg}}^2$$

Using a Taylor series expansion, the covariance between births in the year t and in the year u, $1 \le t \le u \le 16$, is given by

$$\text{cov}\left(\tilde{p}(0,t), \tilde{p}(0,u)\right) \approx \left(1/\text{P}(0,t)\right)\left(1/\text{P}(0,u)\right)\sum_{j=15}^{44}\sum_{k=15}^{44}\text{B}(j,t)\text{B}(k,u)\{{\sigma_{JO}}^2 + t{\sigma_{ft}}^2 + \text{cov}(V(t),V(u))\}$$

the covariance between $V(t)$ and $V(u)$ is given by

$$\begin{aligned}
\text{cov}(V(t),V(u)) &= \text{cov}\left[\sum_{k=0}^{t-1}\left(\varepsilon_{sv}(k)+\varepsilon_{mg}(k)\right), \sum_{h=0}^{u-1}\left(\varepsilon_{sv}(h)+\varepsilon_{mg}(h)\right),\right] \\
&= \text{cov}\left(\sum_{k=0}^{t-1}\varepsilon_{mg}(k), \sum_{h=0}^{u-1}\varepsilon_{mg}(h)\right) + \text{cov}\left(\sum_{k=0}^{t-1}\varepsilon_{sv}(k), \sum_{h=0}^{u-1}\varepsilon_{sv}(h)\right) \\
&= t{\sigma_{mg}}^2 + \text{cov}\left(\sum_{a=0}^{t-1}(t-a)e_{sv}(a), \sum_{b=0}^{u-1}(u-b)e_{sv}(b),\right) \\
&= t{\sigma_{mg}}^2 + \sum_{a=0}^{t-1}\sum_{b=0}^{u-1}(t-a)(u-b)\,\text{cov}(e_{sv}(a), e_{sv}(b)) \\
&= t{\sigma_{mg}}^2 + {\sigma_{sv}}^2\sum_{a=0}^{t-1}(t-a)(u-a) \\
&= t{\sigma_{mg}}^2 + {\sigma_{sv}}^2(u+t+1)(t+1)t/6
\end{aligned}$$

It follows that

$$\text{cov}\left(\tilde{p}(0,t), \tilde{p}(0,u)\right) \approx {\sigma_{JO}}^2 + t{\sigma_{ft}}^2 + t{\sigma_{mg}}^2 + {\sigma_{sv}}^2(u+t+1)(t+1)t/6,$$

and

$$\text{var}\left(\tilde{p}(0,t)\right) \approx {\sigma_{JO}}^2 + t{\sigma_{ft}}^2 + t{\sigma_{mg}}^2 + {\sigma_{sv}}^2 g(t)$$

First, for $1 \le t \le 16 \le u \le 32$,

Next, Alho considers the second generation of births and their survival. That is when the births generated during first 16 years contribute new births, $17 \le t \le 32$. The contribution of the jump-off population, mortality and migration to the uncertainty of $\tilde{p}(0,t)$ is $e_{JO} + V(t)$.

The contribution of fertility, however, consists of two parts. First, there is the direct contribution of fertility. Second, the women of the child bearing ages who were born after the jump-off year contribute to the uncertainty of fertility. This contribution is given by

$$H(t) = \left(1 \Big/ \mathrm{P}(0,t) \right) \sum_{j=15}^{t-2} \mathrm{B}(j,t) \varepsilon_{ft}(t-j-1)$$

Alho considers two types of covariances. First, for $1 \le t \le 16 \le u \le 32$,

$$\begin{aligned} \mathrm{cov}\left(\tilde{p}(0,t), \tilde{p}(u) \right) &= \mathrm{cov}\left[e_{JO} + V(t) + \varepsilon_f(t), e_{JO} + V(u) + \varepsilon_{ft}(u) + H(u) \right] \\ &= {\sigma_{JO}}^2 + t{\sigma_{mg}}^2 + {\sigma_{sv}}^2 (u+t+1)(t+1)t/6 + t{\sigma_{ft}}^2 + \\ &\quad \mathrm{cov}\left(\varepsilon_{ft}(t), H(u) \right) \end{aligned}$$

The last covariance term is given by

$$\begin{aligned} \mathrm{cov}\left(\varepsilon_{ft}(t), H(u) \right) &= 1 \Big/ \mathrm{P}(0,t) \sum_{j=15}^{u-2} \mathrm{B}(j,u) \left[\mathrm{cov}\left(\varepsilon_{ft}(t), \varepsilon_{ft}(u-1-j) \right) \right] \\ &= 1 \Big/ \mathrm{P}(0,u) \sum_{j=15}^{u-2} \mathrm{B}(j,u) {\sigma_{ft}}^2 \min\{t, u-1-j\} \end{aligned}$$

And for $17 \le t \le u \le 32$ this covariance is given by

$$\begin{aligned} \mathrm{cov}\left(\tilde{p}(0,t), \tilde{p}(u) \right) &= \mathrm{cov}\left[e_{JO} + V(t) + \varepsilon_{ft}(t) + H(t), e_{JO} + V(u) + \varepsilon_{ft}(u) + H(u) \right] \\ &= {\sigma_{JO}}^2 + t{\sigma_{mg}}^2 + {\sigma_{sv}}^2 (u+t+1)(t+1)t/6 + t{\sigma_{ft}}^2 + \mathrm{cov}\left[\varepsilon_{ft}(t), H(u) \right] + \\ &\quad \mathrm{cov}\left[H(t), \varepsilon_{ft}(u) \right] + \mathrm{cov}\left[H(t), H(u) \right] \end{aligned}$$

The covariance terms are given as follows

$$\mathrm{cov}\left[\varepsilon_{ft}(t), H(u) \right] = 1 \Big/ \mathrm{P}(0,u) \sum_{j=15}^{u-2} \mathrm{B}(j,u) {\sigma_{ft}}^2 (u-1-j),$$

$$\mathrm{cov}\left[H(t), \varepsilon_{ft}(u) \right] = 1 \Big/ \mathrm{P}(0,t) \sum_{j=15}^{t-2} \mathrm{B}(j,t) {\sigma_{ft}}^2 (t-1-j), \text{ and}$$

$$\mathrm{cov}\left[H(t), H(u) \right] = 1 \Big/ \mathrm{P}(0,t) \, 1 \Big/ \mathrm{P}(0,u) \sum_{j=15}^{t-2} \sum_{k=15}^{u-2} \mathrm{B}(j,t) \mathrm{B}(k,u) {\sigma_{ft}}^2 \min\{t-1-j, u-1-k\}$$

Alho considers the third generation of births, i.e. the years $t = 33,...,48$. Again the authors splits the contribution of fertility into direct and indirect parts. The direct part is $\varepsilon_{ft}(t)$. The indirect part is given by

$$D'(t) = {}^{1}\!/\!_{\mathrm{P}(0,t)} \sum_{j=15}^{\min\{44,t-2\}} \mathrm{B}(j,t) H'(t-1-j)$$

If $t-1-j \leq 16$ then $H'(t-1-j) = \varepsilon_{ft}(t-1-j)$, and if $t-1-j > 16$, then $H'(t-1-j) = \varepsilon_{ft}(t-1-j) + H(t-1-j)$. For $t-1-j \leq 16$ the covariance is given by

$$\mathrm{cov}\left[H'(t-1-j), \varepsilon_{ft}(t)\right] = \sigma_{ft}^{\,2}(t-1-j)$$

For $t-1-j > 16$, the covariance is given by

$$\begin{aligned}
\mathrm{cov}\left[H'(t-1-j), \varepsilon_{ft}(t)\right] &= \mathrm{cov}\left[\varepsilon_{ft}(t-1-j) + H(t-1-j), \varepsilon_{ft}(t)\right] \\
&= \sigma_{ft}(t-1-j) + \mathrm{cov}\left[H(t-1-j), \varepsilon_{ft}(t)\right] \\
&= \sigma_{ft}(t-1-j) + \\
&\quad {}^{1}\!/\!_{\mathrm{P}(0,t-1-j)} \sum_{k=15}^{t-3-j} \mathrm{B}(k,t-1-j)\sigma_{ft}^{\,2}(t-1-j)
\end{aligned}$$

A.5 Theorems of population projection

This section discusses theorems of population projection. First, theorems developed by Cohen (1977) are described. Second, the theorems of Heyde and Cohen (1985) are viewed. This is followed by a discussion of results presented in Tuljapurkar (1990).

Cohen (1977) developed the ergodic theorems of demography. Given certain assumptions about the projection matrix, $\mathrm{A}(t)$, ergodic theorems describe the long run behaviour of population size, $\mathrm{P}(t)$, and of age structure, $q(t)$, and show that the behaviour of these quantities is independent of the initial conditions.

Cohen starts with a set of assumptions. First, the author assumes a finite number of age classes, l. Second, he considers a population subjected only to birth and death, with no immigration or emigration. Third, only one sex is considered, human females, and the vital rates refer to birth and death rates. It is further assumed that the age-specific vital rates apply to all individuals in an age class uniformly and equally. Finally, Cohen considers only large populations.

The author presents next a set of definitions. Cohen defines $\mathrm{P}(t)$ as a non-negative vector representing the age census at time t, with $\mathrm{P}(j,t)$ representing the number of females at time t who will be j years old at their next birthday. The age structure, $q(t)$, of an age census $\mathrm{P}(t)$ is given by $\mathrm{P}(t)/\|\mathrm{P}(t)\|$, with $\|q(t)\| = 1$. Furthermore, $\mathrm{A}(t)$ is defined as a sequence of operators mapping the non-negative $l-$vectors at one time to the non-negative vectors at the next time. In other words, Cohen considers the model

$$\mathrm{P}(t+1) = \mathrm{A}(t+1)\mathrm{P}(t),\ t = 0,1,... \qquad \text{(A.5.1)}$$

More assumptions follow. The author assumes that each $\mathrm{A}(t)$ is a linear operator, represented by a $l \times l$ matrix, where the ratio of the smallest positive element of $\mathrm{A}(t)$ to the largest element of $\mathrm{A}(t)$ is not less than $\nu > 0$. This means that the set of all projection matrices, C, is an ergodic set of matrices. In addition, every element of A^r is positive, i.e. every product of every r matrices from C is positive.

Cohen presents next the strong ergodic theorem. Letting A be a $l \times l$ primitive matrix, then the eigenvalue of A, λ, has an algebraic multiplicity one, and geometric multiplicity one. Furthermore, λ is real and positive. Defining v_1 and v_2 as the left and right eigenvector respectively, then $\lim_{t\to\infty}\left(\mathrm{A}(t)/\lambda^t\right) = v'_1 v_2$, where v_1 and v_2 are scaled so that $v'_2 v_1 = 1$. Letting $\mathrm{A}(t) = \mathrm{A} \in C$, and $\mathrm{P}(0) \neq 0, \mathrm{P}'(0) \neq 0$, $\mathrm{P}(0) \neq \mathrm{P}'(0)$, be two non-negative non-zero and different age censuses, this leads to $\mathrm{P}(t) = \mathrm{A}^t\mathrm{P}(0)$ and $\mathrm{P}'(t) = \mathrm{A}^t\mathrm{P}'(0)$. Then

$$\lim_{t\to\infty} \mathrm{P}(t)/\lambda^t = v_1\left(v'_2\mathrm{P}(0)\right)$$

In addition

$$\lim_{t\to\infty} q(t) = \lim_{t\to\infty} q'(t) = v_1 / \| v_1 \|$$

The author makes the following remarks. The eigenvalue λ is called the stable growth rate per unit of time. The Malthusian parameter or intrinsic rate of natural increase is defined by $\log(\lambda)$. The stable age-structure is defined by $v_1 / \| v_2 \|$. The strong ergodic theorem asserts that $\mathrm{P}(t)$ and $\mathrm{P}'(t)$ grow at the same rate, and that the age structures $q(t)$ and $q'(t)$ approach the same limiting age structure.

Next, the author presents the weak ergodic theorem. If $\mathrm{A}(1), \mathrm{A}(2),...$ are projection matrices, repetition possible, and $\mathrm{P}(0)$ and $\mathrm{P}'(0)$ are two different initial age censuses, implying that

$$\mathrm{P}(t) = \mathrm{A}(t),..., \mathrm{A}(1)\mathrm{P}(0), \text{ and}$$

$$\mathrm{P}'(t) = \mathrm{A}(t),..., \mathrm{A}(1)\mathrm{P}'(0)$$

Then

$$\lim_{t\to\infty} \| q(t) - q'(t) \| = 0$$

That means that regardless of the initial age-structure the vital rates in the matrix A determine the current age-structure.

The weak stochastic ergodic theorem follows. If the sequence of projection matrices applied to $\mathrm{P}(0)$ is a sample path of a Markov chain, then the joint process consisting of $\mathrm{A}(t)$ and $q(t)$ is a Markov chain with transition function $G(t)$ given by

$$G(t) = \Pr\{\mathrm{A}(t+1) \in C, q(t+1) \in D \mid \mathrm{A}(t), q(t)\}$$

Suppose that the projection matrices are chosen from an ergodic set of projection matrices, and that the Markov chain is S-uniformly ergodic . Then the Markov chain $(\mathrm{A}(t), q(t))$ is uniformly weakly ergodic in the sense that for every origin of time, and for every $\delta' > 0$, and for every measurable set C of projection matrices and every measurable set D of age-structures, there exists an integer a_0 a such that for all $a > a_0$

$$\sup_{(A,q),(A',q')} |\Pr[(\mathrm{A}(a), q(a)) \in (C, D) \mid \mathrm{A}(1), q(1) = (\mathrm{A}, q)] - \Pr[(\mathrm{A}(a), q(a)) \in (C, D) \mid \mathrm{A}(1), q(1) = (\mathrm{A}', q')]| < \delta'$$

This means that the joint distribution of the current projection matrix and current age structure, $(\mathrm{A}(t), q(t))$ becomes independent of the initial projection matrix and initial age-structure after a long time, uniformly with respect to initial conditions.

Finally, Cohen discusses the strong stochastic ergodic theorem. This theorem can be translated into practice as follows. When the set C contains a finite number of projection matrices, and the Markov chain on C is homogenous and regular, the long run rate of growth of the expected population size is the dominant eigenvalue of a certain matrix. The long run age-structure of the expected population maybe calculated from the dominant eigenvector of this matrix. Technically, the theory states that when the Markov chain on A is homogenous, i.e. when the probability of transition from one projection matrix to another is constant in time, the joint distribution $F(t)$ of the current $(\mathrm{A}(t), q(t))$ approaches a limiting invariant probability distribution, F , which is the solution of

$$F(C, d) = \int F(d\mathrm{A}, dq)(\mathrm{A}, q, C, D)$$

Equipped with these theorems, Cohen suggests a scheme for using historical data in population projection. The scheme starts with arranging all age-specific effective fertility and survival rates in a projection matrix into a vector. Then a linear first order auto-regressive model (see Appendix B section B.1.2) should be fitted to historically observed sequences of such vectors. The initial vital rates and the estimated parameters are used to project future vital rates. The distribution of future vital rates, given an initial age-structure, implies a distribution of the projected subsequent age-structure and population sizes.

Heyde and Cohen (1985) also established three theorems for population projections. The analysis is based on vital rates that vary stochastically in time. The authors consider the model

$$\mathrm{P}_\omega(t+1) = \mathrm{A}_\omega(t+1)\mathrm{P}_\omega(t),$$

where $\mathrm{P}(t)$ is a vector of the number of individuals in each age class at time t, A is the matrix of vital rates, and ω refers to a particular realization of the process that produces the vital rates.

Heyde and Cohen proceed with a set of definitions and assumptions. Letting $\mathrm{A}(1), \mathrm{A}(2), \ldots$ be a stationary ergodic random sequence of $l \times l$ matrices with non-negative elements, this sequence of matrices satisfies two conditions. First, there exists an integer a such that any product

$\mathrm{A}(i+a)...\mathrm{A}(i+1)$ of the matrices has all its elements positive with probability one. Second, for another constant b, $1 < b < \infty$, and each matrix $\mathrm{A}(i)$, it follows that

$$1 \le \max(\mathrm{A}(i)) / \min(\mathrm{A}(i)) \le b$$

Furthermore, The authors define the concept of uniform mixing as follows. Assuming that $\mathrm{A}(1), \mathrm{A}(2),...$ are defined on a probability space $[\Omega, \wp, P]$ (Billingsley 1995, pp. 23), and $\wp_t^u$ is the $\sigma - field$ (Billingsley 1995, pp. 20) generated by $\mathrm{A}(t)...\mathrm{A}(u)$, and letting

$$\varphi(n) = \sup(|\Pr(Event\ 2 \mid Event\ 1) - \Pr(Event\ 2)|; Event\ 1 \in \wp_o^i, Event\ 2 \in \wp_{i+n}^{\infty}, \Pr(Event\ 1) > 0),$$

for two events $Event\ 1$ and $Event\ 2$. Then the condition $\varphi(n) \to 0$ as $n \to \infty$ is called uniform mixing.

Next, the authors present the first theorem. Supposing that the stationary ergodic sequence of matrices satisfies the two conditions mentioned above, and that $E|\log(\max(\mathrm{A}(1))| < \infty$, then for all $1 \le i, j \le k$

$$t^{-1}\log(\mathrm{A}(t)...\mathrm{A}(1))_{ij} \xrightarrow{a.s.} \log(\lambda), \tag{A.5.2}$$

where λ is a finite constant as $t \to \infty$. In addition, if

$$E|\log(\mathrm{A}(1))|^2 < \infty, \tag{A.5.3}$$

and

$$\sum_{n=1}^{\infty} |\varphi(n)|^{1/2} < \infty \tag{A.5.4}$$

then

$$\lim_{t\to\infty} t^{-1/2} E|\log(\mathrm{A}(t)...\mathrm{A}(1))_{ij} - t\log\lambda| = \sigma(2\pi)^{1/2}$$

exists for $0 \le \sigma \le \infty$. If $\sigma > 0$, then

$$(t\sigma^2)^{-1/2}[\log(\mathrm{A}(t)...(1))_{ij} - t\log\lambda] \xrightarrow{d} \mathrm{N}(0,1),$$

as $t \to \infty$.

Heyde and Cohen proceed with the second theorem. Letting $\mathrm{P}(t+1) = \mathrm{A}(t+1)\mathrm{P}(t)$, $t \ge 0$, and $W'(t) = (a', \mathrm{P}(t))$, where a' is a non-zero vector of non-negative elements, then

$$\lim_{t\to} t^{-1}\log W'(t) = \log(\lambda) \text{ a.s.}$$

and if (A.5.3) and (A.5.4) hold then

$$\lim_{t\to\infty} t^{-1/2} E \mid \log\left(W^{'}(t)\right) - t\log(\lambda) \mid = \sigma(2\pi)^{1/2}$$

exists for $0 \le \sigma \le \infty$, and if $\sigma > 0$, then

$$\left(t\sigma^2\right)^{-1/2}\left\{\log\left(W^{'}(t)\right) - t\log(\lambda)\right\} \xrightarrow{d} \mathrm{N}(0,1),$$

as $t \to \infty$.
Finally, the authors present a third theorem. If (A.5.1), (A.5.2) and (A.5.3) hold, then

$$\left(\log t\right)^{-1} \sum_{i=1}^{t} \mid \log\left(W^{'}(i)\right) - i\log\left(\hat{\lambda}\right) \mid i^{-3/2} \xrightarrow{p} \sigma(2\pi)^{1/2},$$

as $t \to \infty$, where $\log(\hat{\lambda}) = t^{-1}\log\left(W^{'}(t)\right)$.
The authors propose confidence intervals for the growth rate and total population size. Usually, $W^{'}(t)$ is the population size, $W(t)$, given by $W(t) = \left(e^{'}, \mathrm{P}(t)\right)$. An approximate $100(1-\alpha)\%$ confidence interval for the growth rate $\log(\lambda)$ is given by

$$t^{-1}\log\left(W(t)\right) \pm z_{\alpha/2}\,\hat{\sigma}\,t^{-1/2}$$

In practice, the actual generation numbers at times $t = 1,...,T$ will usually be unknown. The difference from $t = 1$ is used instead. This results in the following formula for $\log(\hat{\lambda})$

$$\log(\hat{\lambda}) = (\mathrm{T}-1)^{-1}\log\left(W(\mathrm{T}) - W(1)\right),$$

and this formula for $\hat{\sigma}$

$$\hat{\sigma} = (\pi/2)^{1/2}\left(\log(\mathrm{T}-1)^{-1}\sum_{i=1}^{\mathrm{T}-1} i^{-3/2} \mid \log\left(W(i+1)\right) - \log\left(W(1)\right) - i\log(\hat{\lambda}) \mid\right.$$

(For the confidence interval for the population size and its derivation see section 3.4.)
Tuljapurkar (1990) added more results to the theories mentioned above. The author developed the new results based on random matrices products, and on random vital rates.
The author starts with a three assumptions. First, the demographic weak ergodicity theory holds. Second, the random process generating vital rates is stationary and ergodic. Third, the logarithmic moment of vital rates is bounded. With $\log_{+}(x) = \max\left\{0, \log(x)\right\}$, and $\|.\|$ any matrix norm, this translates to

$$E\log_{+} \| \mathrm{A}(1) \| < \infty$$

This set of three assumptions is to be called assumptions set one.
The first result concerns the long run growth rate. The long run growth rate of the log of total population, or any part of the population, is almost surely given by a number, $\log(\lambda)$, independent of the initial population vector. This number is given by

$$\log(\lambda) = \lim_{t\to\infty}\left\{\log\left(a^{''}, \mathrm{P}(t)\right)/t\right\},$$

where $a^{''}$ is a vector of bounded non-negative numbers.
Tuljapurkar's next result concerns age-structure. Starting from every initial age-structure, $q(0)$, the population converges to a time dependent stationary random sequence of structure vectors, $\hat{q}(t)$, which are independent of $q(0)$. There is a stationary measure describing the probability distribution of the joint sequence of vital rates and population structure vectors $\{\mathrm{A}(1), q(1), \mathrm{A}(2), q(2)...\}$.

The following result concerns the growth rate. There are constants c_i, $i = 1,..., k$, such that

$$\log(\lambda) = c_1 \geq c_2 \geq ...$$

The constants c_i are determined by growth rates the exterior powers of the A's (Lang 1984).
Before proceeding with the next result, the author makes a new assumption. It is assumed that the random process generating vital rates can be run backwards in time, creating a unique time-reversed process that is stationary and ergodic. Tuljapurkar then considers the adjoint time-reversed process associated with

$$Q^{'}(t) = \mathrm{A}^{'}(t)Q(t+1)/\left(e^{'}, \mathrm{A}^{'}(t)Q^{'}(t+1)\right) \qquad \text{(A.5.5)}$$

Then A.5.5 runs backwards in time through decreasing values of t, and as $t \to -\infty$, the resulting vectors $Q^{'}(t)$ converge to a stationary random sequence of vectors $\hat{Q}^{'}(t)$.
The following result is about the asymptotic distribution of total population size. With assumptions set one still holding, it is further assumed that the random process generating vital rates is rapidly mixing (king 2003)**,** to be called assumption two here. Then the total population size at time t, $W(t) = \left(e^{'}, \mathrm{P}(t)\right)$, is asymptotically distributed as log normal. In other words

$$\log\left\{\left(W(t) - t\log(\lambda)\right)/\left(\sigma/\sqrt{t}\right)\right\} \to \mathrm{N}(0,1),$$

for some σ. (Tuljapurkar refers to Hedye and Cohen (1985) for methods for estimation of σ.)
The author proceeds with the joint distribution of the vital rates and population structure. First, assumptions set one and assumption two still hold. Second, it is assumed that the vital rates follow a countable state Markov process. Then, there is a joint probability distribution of vital rates and population structures given by

$$F(t, C, D) = \Pr\{\mathrm{A}(t) \in C, q(t) \in D\}$$

As $t \to \infty$, F converges to an equilibrium distribution, say $\overset{*}{F}(C, D)$.
Equation A.5.1 can be written in terms of the age structure in the following way

$$q(t+1) = \frac{\mathrm{A}(t+1)q(t)}{\left(e', \mathrm{A}(t+1)q(t)\right)},$$

using the fact that $\frac{W(t+1)}{W(t)} = \left(e', \mathrm{A}(t+1)q(t)\right)$. Tuljapurkar (1990) argues that the average growth rate can be computed as the average one-time step growth rate given by

$$\log(\lambda) = E\log\left(e', \mathrm{A}(1)q(0)\right)$$

Tuljapurkar considers next the moments of the population vector. It is first assumed that the vital rates follow a finite state Markov process. Then the moments of the population vector and its tensor powers, $E\left(\mathrm{P}(t) \otimes \mathrm{P}(t)\right), E\left(\mathrm{P}(t) \otimes \mathrm{P}(t) \otimes \mathrm{P}(t)\right), \ldots$, can be computed explicitly as a function of time. Given the basic model $\mathrm{P}(t+1) = \mathrm{A}(t+1)\mathrm{P}(t)$, the average population vector is given by

$$E\left(\mathrm{P}(t)\right) = E\left(\mathrm{A}(t)\mathrm{P}(t-1)\right) = E\left(\mathrm{A}(t)\right)E\left(\mathrm{P}(t-1)\right),$$

where $\mathrm{P}(t)$ is independent of the history of the population. The second moment $E\left(\mathrm{P}(t+1) \otimes \mathrm{P}(t+1)\right)$ is given by

$$\begin{aligned} E\left(\mathrm{P}(t+1) \otimes \mathrm{P}(t+1)\right) &= E\left(\mathrm{A}(t+1)\mathrm{P}(t)\right) \otimes \left(\mathrm{A}(t+1)\mathrm{P}(t)\right) \\ &= E\left(\mathrm{A}(t+1) \otimes \mathrm{A}(t+1)\right)\left(\mathrm{P}(t) \otimes \mathrm{P}(t)\right) \\ &= E\left(\mathrm{A}(t+1) \otimes \mathrm{A}(t+1)\right)E\left(\mathrm{P}(t) \otimes \mathrm{P}(t)\right) \end{aligned}$$

Higher moments are driven in an analogous manner.
Tuljapurkar's final result concerns the probability distribution of the age-structure. One more assumption to be added is that the I.I.D. model determines the random vital rates. (Under the I.I.D. model the entries of the $\mathrm{A}(t)$'s are chosen randomly for each t from the same fixed distribution. There is no serial correlation between vital rates at different times, but rates within each period can be correlated. The number of possible environments can be finite or infinite, with the environment being totally unpredictable.) Then, there is a probability distribution for the population structure vector given by

$$G(t, D) = \Pr\left(q(t) \in D\right),$$

and a corresponding stationary distribution, $\overset{*}{G}(D)$, to which $G(t, D)$ converges as t increases.
Armed with these theorems, Tuljapurkar suggests the following method for population projection. The random rates model leads asymptotically to simple exponential growth model described by the lognormal theorem. If the assumptions are satisfied for a set of historical data, projections can be made by estimating the parameters $\log(\lambda)$ and σ. For $\log(\lambda)$ Tuljapurkar suggests using

$$\log\left(\hat{\lambda}\right) = \{\log(W(T) - W(1))\}/(T-1)$$

Appendix B

B.1 Time series analysis

B.1.1 General theory

A time series is a stochastic process where the time index takes on a finite or countably infinite set of values. The general expression for a time series is given by

$$y(t) = f\left(y(t-1), y(t-2), ..., \varepsilon(t)\right),$$

where $\varepsilon(t)$ is a disturbance term. The functional form f, number of lags and a structure for the disturbance term must be specified. The functional form can be the general p^{th} order auto-regressive process expressed by

$$y(t) = \alpha_1 y(t-1) + ... + \alpha_p y(t-p) + \varepsilon(t)$$

If $\varepsilon(t)$ is assumed to be white noise, then the process is called a pure AR(p) process. The disturbance term is said to be white noise if

$$\begin{aligned} E\left(\varepsilon(t)\right) &= 0 \\ E\left(\varepsilon^2(t)\right) &= \sigma^2, \text{ for all t} \\ E\left(\varepsilon'(t)\varepsilon(u)\right) &= 0, \text{ for t} \neq \text{u} \end{aligned}$$

When the disturbance term is not assumed to be white noise, its usual specification is a moving average, MA(q), process given by

$$\varepsilon(t) = \varepsilon'(t) - \beta_1 \varepsilon'(t-1) - ... - \beta_q \varepsilon'(t-q),$$

where ε' is a white noise process. The MA(q) process assumes a more complicated structure for the disturbance term in the AR(p) process. The AR(p) process and the MA(q) process can be combined to form a mixed auto-regressive moving average ARMA(p,q) process given by

$$y(t) = \alpha_1 y(t-1) + ... + \alpha_p y(t-p) + \varepsilon'(t) - \beta_1 \varepsilon'(t-1) - ... - \beta_q \varepsilon'(t-q)$$

There are three steps in ARMA modelling. First, the series should be checked for stationarity. Second, for purposes of estimation and testing an ARMA specification should be chosen. Third, from the preferred specification forecasts are calculated over a relevant time horizon.
The process starts with checking stationarity. A stationary series has a constant unconditional mean at all points and a constant unconditional variance independent of time. Stationarity can be checked using unit root tests, e.g. the Dickey-Fuller test. If the series is found to be non-stationary, it needs to be first differenced to yield a stationary one. The minimum number of time the series needs to be first differenced to give a stationary series is called the order of integration. An ARMA(p,q) which has order of integration u is denoted ARIMA(p,u,q).
The next step is model specification. Choosing the order of p and q can be done using the method of Hannan and Rissanen (Johnston and Dinardo1997).

Estimating the parameters of the model follows. This can be done using Least Squares or Maximum Likelihood methods. Two conditions must be satisfied though. First, the disturbances must be independently and identically distributed. Second, the series must be stationary. For MA models non-linear methods are called for.
Once the parameters are estimated, forecasting can be carried out. Given that the observations on y are available for periods 1 to T, forecasts are made based on information at time T. Let

$$y(T+s) = \text{value of } y \text{ at period } T+s,\ s>0,$$

$$\hat{y}(T+s) = \text{forecasts of } y(T+s) \text{ based on information available at time } T, \text{ and}$$

$$e(T+s) = y(T+s) - \hat{y}(T+s)$$

The forecast of $y(T+s)$ with the minimum mean squared error is the conditional expectation of $y(T+s)$, given information available at time T (Hamilton 1994, pp. 73).

B.1.2 The AR(1) process

The AR(1) process is given by

$$y(t) = \alpha_0 + \alpha_1 y(t-1) + \varepsilon(t),$$

where $\varepsilon(t)$ is white noise. Using the lag operator (Johnston and DiNardo 1997, pp. 206) the process can be written as

$$y(t) = \alpha_0(1+\alpha_1+\alpha_1^{\,2}+\ldots) + (\varepsilon(t)+\alpha_1\varepsilon(t-1)+\alpha_1^{\,2}(\varepsilon(t-2)+\ldots)$$

It follows that

$$E\big(y(t)\big) = \alpha_0(1+\alpha_1+\alpha_1^{\,2}+\ldots) = \frac{\alpha_0}{1-\alpha_1}$$

The necessary and sufficient condition for the existence of the expectation is $\alpha_1 < 1$, so that the y series has a constant and unconditional mean independent of time, therefore stationary. For the variance of $y(t)$ consider

$$y(t) - E\big(y(t)\big) = \varepsilon(t) + \alpha_1\varepsilon(t-1) + \alpha_1^{\,2}\varepsilon(t-2) + \ldots$$

This indicates that the variance of $y(t)$ is given by

$$\text{var}\big(y(t)\big) = \sigma_y^{\,2} = \sigma_\varepsilon^{\,2} + \alpha_1^{\,2}\sigma_\varepsilon^{\,2} + \alpha^4{}_1\sigma_\varepsilon^{\,4} + \ldots = \frac{\sigma_\varepsilon^{\,2}}{1-\alpha_1^{\,2}}$$

The variance of $y(t)$ is constant, unconditional and independent of time. The autocovariance is given by

$$\gamma_1 = E\big(y(t) - E(y(t))\big)\big(y(t-1) - E(y(t))\big) = \alpha_1 \frac{\sigma_\varepsilon^{\,2}}{1-\alpha_1^{\,2}},$$

and

$$\gamma_r = E\big(y(t) - E(y(t))\big)\big(y(t-r) - E(y(t))\big) = \alpha_1^{\,r} \frac{\sigma_\varepsilon^{\,2}}{1-\alpha_1^{\,2}}, \quad r = 0,1,2,...$$

The mean, variance and covariance are all constants independent of time. The autocorrelation coefficients are given by

$$\rho_r = \frac{\gamma_r}{\gamma_0} = \frac{\gamma_r}{\sigma_y^{\,2}} = \left(\frac{E\big(y(t)y(t-r)\big)}{\sqrt{\mathrm{var}\big(y(t)\big)}\sqrt{\mathrm{var}\big(y(t-r)\big)}} \right)$$

The estimation is done by OLS. The values of $y(1)$ are taken as given and summation run over $t = 2,3,...,n$. For forecasting write the AR(1) process as follows

$$y(t) - E\big(y(t)\big) = \alpha_0 + \alpha_1 y(t-1) + \varepsilon(t) - \frac{\alpha_0}{1-\alpha_1} = \alpha_1 \left(y(t-1) - \frac{\alpha_0}{1-\alpha_1} \right) + \varepsilon(t),$$

with $\alpha_1 < 1$, and $\varepsilon(t)$ *i.i.d.* with mean zero and variance $\sigma_\varepsilon^{\,2}$. Or

$$y(t) = \frac{\alpha_0}{1-\alpha_1}(1-\alpha_1) + \alpha_1 y(t-1) + \varepsilon(t)$$

As mentioned before the forecast minimizing the MSE is the conditional expectation of $y(T+s)$, given information available at time T. If it is assumed that observations on y are available for periods 1 to T, then the forecasts are made conditional on information available at time T. Thus

$$\begin{aligned} \hat{y}(T+1) &= E\big(y(T+1) \mid y(T)\big) \\ &= E\left\{ \frac{\alpha_0}{1-\alpha_1}(1-\alpha_1) + \alpha_1 y(T) + \varepsilon(T+1) \right\} \mid y(T) \\ &= (1-\alpha_1)\frac{\alpha_0}{1-\alpha_1} + \alpha_1 y(T), \end{aligned}$$

and $e(T+1) = \varepsilon(T+1)$, indicating that $\mathrm{var}\big(e(T+1)\big) = \sigma_\varepsilon^{\,2}$. Similarly, $y(T+2)$ can be written in the following way

$$y(T+2) = (1-\alpha_1)\frac{\alpha_0}{1-\alpha_1} + \alpha_1{}^2 y(T) + \alpha_1 \varepsilon(T+1) + \varepsilon(T+2)$$

Then

$$\hat{y}(T+2) = (1-\alpha_1{}^2)\frac{\alpha_0}{1-\alpha_1} + \alpha_1{}^2 y(T),$$

$$e(T+2) = \alpha_1 \varepsilon(T+1) + \varepsilon(T+2), \text{ and}$$

$$\text{var}\Big(e(T+2)\Big) = \sigma_\varepsilon{}^2\left(1+\alpha^2{}_1\right)$$

Proceeding the same way $y(T+s)$ can be written as

$$y(T+s) = (1-\alpha^s{}_1)\frac{\alpha_0}{1-\alpha_1} + \alpha_1{}^s y(T) + \left(\varepsilon(T+s) + \alpha_1 \varepsilon(T+s-1) + ... + \alpha^{s-1}\varepsilon(T+1)\right)$$

,

and

$$\hat{y}(T+s) - \frac{\alpha_0}{1-\alpha_1} = \alpha^s\left(y(T) - \frac{\alpha_0}{1-\alpha_1}\right)$$

and the forecast error variance is given by

$$\text{var}\big(e(T+s)\big) = 1 + \alpha_1{}^2 + \alpha_1{}^4 + ... + \alpha_1{}^{2(s-1)}\sigma_\varepsilon{}^2$$

As $s \to \infty$, $\hat{y}(T+s) \to \frac{\alpha_0}{1-\alpha_1}$, and $\text{var}\big(e(T+s)\big) \to \frac{\sigma_\varepsilon{}^2}{1-\alpha^2}$. Therefore, as the forecast horizon increases, the forecast value tends to the unconditional mean of the process, and the forecast error variance tends to the unconditional variance of the process.

B.1.3 Heteroscedacity and autocorrelation

The aforementioned procedures are no more valid if the white noise assumption is violated which can be either due to heteroscedacity or autocorrelation. In the case of heteroscedacity Generalized Least Squares (GLS) can be applied (Johnston and DiNardo 1997, pp. 170). When the disturbances are autocorrelated, which can be a sign of incorrect specification of the model, the assumption of zero pair wise covariance does not hold.
There are several tests for autocorrelation (Johnston and DiNardo 1997), but here the Durbin Watson test is discussed. Consider the following AR(1) formulation for the autocorrelation

$$\varepsilon(t) = \alpha\varepsilon(t-1) + \xi(t),$$

where $\xi(t)$ is white noise. The null hypothesis of zero correlation $H_0 : \alpha = 0$ is tested against the alternative hypothesis $H_1 : \alpha \neq 0$. The Durbin Watson statistic is computed from the vector of Ordinary Least Squares (OLS) residuals $\mathrm{Y} - \mathrm{X}\beta$. The test statistic is given by

$$DW = \frac{\sum_{t=2}^{n} \left(e(t) - e(t-1)\right)^2}{\sum_{t=1}^{n} e^2(t)}$$

For large n this reduces to $DW \approx 2(1-\hat{\alpha})$, where

$$\hat{\alpha} = \frac{\sum_{t=2}^{n} \left(e(t)e(t-1)\right)}{\sum_{t=2}^{n} e^2(t-1)},$$

is the coefficient in the OLS regression of $e(t)$ on $e(t-1)$. Ignoring end-point discrepancies $\hat{\alpha}$ can be seen as an approximation to the simple correlation coefficient between $e(t)$ on $e(t-1)$. That indicates that the value of DW will be less than two for positive autocorrelation, greater than two for negative autocorrelation, and approximately two for zero correlation. To test the hypothesis of zero autocorrelation against the alternative of positive first order autocorrelation, Durbin and Watson established upper and lower bounds for the critical values (Johnston en DiNardo 1997, pp. 181).
In order to apply the Durbin-Watson test two conditions must be satisfied. First, the regression must include a constant term. Second, it is only valid for non-stochastic X matrix, i.e. not when lagged values of the dependent variable are among the regressors. Durbin, however, derived a test for the case when lagged values of the dependent variable are among the regressors (Johnston en DiNardo 1997, pp. 182).
Once a autocorrelation is suggested, estimation follows. The most common specification of autocorrelation is that of a first order autoregressive process. The AR(1) process is given by

$$\varepsilon(t) = \alpha\varepsilon(t-1) + \xi(t),$$

where $\xi(t)$ is white noise. The necessary conditions for stationarity of the AR(1) are $|\alpha| < 1$, $E(\xi(t)) = 0$, $\mathrm{var}(\xi(t) = \sigma_{\varepsilon}^{\ 2}$, and $\sigma^2{}_{\varepsilon} = \sigma_{\xi}^{\ 2} / 1 - \alpha^2$. The autocorrelation coefficients, ρ_r, are given by

$$\rho_r = \alpha^r, \quad r = 0,1,2...$$

The variance-covariance matrix of $\varepsilon(t)$ is given by

$$\operatorname{var}(\varepsilon(t)) = \sigma_\varepsilon{}^2 \begin{bmatrix} 1 & \alpha & . & . & \alpha^{T-1} \\ \alpha & 1 & . & . & \alpha^{T-2} \\ . & . & . & . & . \\ . & . & . & . & . \\ \alpha^{T-1} & \alpha^{T-2} & . & . & 1 \end{bmatrix}$$

$$= \frac{\sigma_\xi{}^2}{1-\alpha^2} \begin{bmatrix} 1 & \alpha & . & . & \alpha^{T-1} \\ \alpha & 1 & . & . & \alpha^{T-2} \\ . & . & . & . & . \\ . & . & . & . & . \\ \alpha^{T-1} & \alpha^{T-2} & . & . & 1 \end{bmatrix} = \sigma_\xi{}^2 \Sigma,$$

with

$$\Sigma = \frac{1}{1-\alpha^2} \begin{bmatrix} 1 & \alpha & . & . & \alpha^{T-1} \\ \alpha & 1 & . & . & \alpha^{T-2} \\ . & . & . & . & . \\ . & . & . & . & . \\ \alpha^{T-1} & \alpha^{T-2} & . & . & 1 \end{bmatrix}$$

If α is known GLS can be applied and the GLS estimate $\hat{\beta}$ of β is given by

$$\hat{\beta}_{GLS} = \left(\mathrm{X}'\Sigma^{-1}\mathrm{X}\right)^{-1}\mathrm{X}'\Sigma^{-1}\mathrm{Y}$$

Alternatively, a matrix Ξ can be found so that $\Sigma^{-1} = \Xi'\Xi$. The data can be transformed and OLS can be applied by regressing $\Xi\mathrm{Y}$ on $\Xi\mathrm{X}$. Usually α is not known and has to be estimated along with the $\hat{\beta}$. In this case Iterative procedures of estimation are called for (Johnston and DiNardo 1997, pp. 191).

B.2 Branching Galton-Watson process

This section discusses the Branching Galton-Watson process. The Branching Galton-Watson process is a class of Markov chain. It originated in 1847 with a mathematical model by Galton and Watson for the problem of extinction of family surnames.

The Branching Galton-Watson process can be defined as follows. Starting with initial set of individuals, P_0, these individuals are called the $zero^{th}$ generation. The offspring produced by the $zero^{th}$ generation forms the first generation, and the offspring produced by the first generation forms the second generation, and so forth. In general, the descendants of the r^{th} generation form the $r+1^{th}$ generation. The number of individuals in the r^{th} generation, $r = 0,1,\ldots$, is a random

variable. Individuals are assumed to reproduce independently of other individuals. Let the probability that an individual produces i similar individuals be ς_i, with $i = 0,1,2,...$, $\sum_{i=0} \varsigma_i = 1$. Then the sequence $\mathrm{P}_0, \mathrm{P}_1, ...$ constitutes a Galton-Watson branching process with offspring distribution ς_i.

A concept associated with the Galton-Watson branching process is that of the probability generating function (PGF). Suppose that P is random variable assuming non-negative integral values $0,1,2,...$ Suppose further that $\Pr(\mathrm{P} = k) = \varsigma_k$, $k = 0,1,2,...$, $\sum_k \varsigma_k = 1$. Then the PGF with a variable r^k, $G(r)$, is defined by

$$G(r) = \sum_{k=0}^{\infty} \varsigma_k r^k = E(r^k)$$

But $E(\mathrm{P}) = \sum_k k\varsigma_k$, and $G'(r) = \sum_{k=0}^{\infty} k\varsigma_k r^{k-1}$. Therefore,

$$E(\mathrm{P}) = \lim_{r \to 1} G'(r) = G'(1)$$

In a similar way it can be seen that $E(\mathrm{P}^2) = G''(1) + G'(1)$, so that the variance of P is given by

$$\mathrm{var}(\mathrm{P}) = G''(1) + G'(1) - (G'(1))^2$$

Note that $\mathrm{P}_n = \sum_{i=1}^{\mathrm{P}_{n-1}} y_i$, where y_i represents the number of offspring of the i^{th} individual of the $(n-1)^{th}$ generation. The y_i's are i.i.d. random variables with distribution ς_i.

Define the P.G.F. $G(r)$ of y_i as $G(r) = \sum_k \Pr(y_i = k) r^k = \sum_k \varsigma_k r^k$, and let $G_n(r) = \sum_k \Pr(\mathrm{P}_n = k) r^k$, $n = 0,1,2,...$ The moments of the branching Galton-Watson process can be found using the P.G.F. The first moment is given by $G'(1) = E(y_1) = E(\mathrm{P}_1) = \mu$. The first moment can be derived as follows

$$\begin{aligned} E(\mathrm{P}_n) &= E[E(\mathrm{P}_n) \mid \mathrm{P}_{n-1}] \\ &= E\left[E \sum_{i=1}^{\mathrm{P}_{n-1}} y_i \mid \mathrm{P}_{n-1} \right] \\ &= E[\mathrm{P}_{n-1}\mu] \\ &= \mu E[\mathrm{P}_{n-1}] \end{aligned}$$

Suppose that $\mathrm{P}_0 = 1$. Then

$$
\begin{aligned}
E(\mathrm{P}_1) &= \mu \\
E(\mathrm{P}_2) &= \mu E(\mathrm{P}_1) = \mu^2 \\
&\vdots \\
E(\mathrm{P}_n) &= \mu^n
\end{aligned}
$$

The variance of the Branching Galton-Watson process is given by (Ross 2000, pp. 204)

$$
Var(\mathrm{P}_n) = \begin{cases} \dfrac{\mu^{n-1}(\mu^n - 1)}{\mu - 1}, \text{if } \mu \neq 1 \\ \mathrm{n}\sigma^2, \text{if } \mu = 1 \end{cases}
$$

When the population consists of a finite number of types of individuals the process is called a multi-type Galton-Watson process. Suppose that a population of individuals originates from a single ancestor and that there are l types of individuals. Let $\varsigma^w(\kappa_1,...,\kappa_l)$ be the probability that an individual of type w, $w = 1,...,l$ produces κ_j offspring of type j, $j = 1,...,l$. The probability generating function of the multi-type Galton-Watson process is given by

$$
G^w(r) = \sum \varsigma^w(\kappa_1,...,\kappa_l) r^{\kappa_1} ... r^{\kappa_l}
$$

Let also $\mathrm{P}_n = \left(\mathrm{P}_n^{(1)},...,\mathrm{P}_n^{(l)}\right)$ represent the population size of l types in the n^{th} generation. The expected number of offspring of type w produced by an individual of type z, u_{wz}, is given by

$$
u_{zw} = \frac{\partial G^z(r)}{\partial r_w}(1,...,1)
$$

B.3 Linear interpolation

A method of estimation and prediction of the value of a variable between two points is linear interpolation. Linear interpolation works by drawing a straight line between two neighbouring samples and returning the appropriate point along that line. Let ϖ be a number between zero and one, representing how far it is intended to interpolate a value y between time t and $t+1$. Then the linearly interpolated value $\hat{y}(t+\varpi)$ is defined by

$$
\hat{y}(t+\varpi) = (1-\varpi)y(t) + \varpi y(t+1)
$$