 doi: 10.1111/j.1467-842X.XXX

Forecasting the age structure of the scientific workforce in Australia

Rob J Hyndman[1] and Khuyen Vanh Nguyen [1]

Monash University

Summary

Planning for a future workforce requires forecasts of age structure changes to inform policy decisions, particularly related to universities and immigration. We propose a new dynamic statistical model for forecasting the age structure of a workforce. Our approach is inspired by a stochastic model used in population forecasting, replacing births with graduate entry, modelling exits through death and retirement, and including a remainder term that captures migration and career changes. Functional data models are used to model age-specific components, while ARIMA models are used for time series components. Simulation is employed to generate forecast distributions, capturing uncertainty from all components. The approach is illustrated using data on Australia's scientific workforce, allowing us to forecast the age distribution of various scientific disciplines for the next ten years. This analysis was central to an Australian Academy of Science initiative examining the capability of Australia's science system and identifying workforce gaps.

Key words: cohort analysis; demographic modelling; functional data models; labour market; workforce planning

1. Introduction

In planning for the future labour market, it is necessary to forecast the age structure of the workforce in order to enable informed decision-making on policies, especially concerning universities and immigration. We propose a statistical modelling approach to this problem, illustrated using various scientific disciplines in Australia, forecasting future workforce age structures over the next decade. The forecasts described have been

[1] Department of Econometrics & Business Statistics, Monash University
Email: `Rob.Hyndman@monash.edu`

Acknowledgment. We thank Alexandra Lucchetti from the Australian Academy of Science for sourcing the data required for this project, and for helpful feedback on earlier versions of this paper. Rob Hyndman is a member of the Australian Research Council Industrial Transformation Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Project ID IC200100009. He also receives funding from the Australian Research Council through Discovery Project DP250100702.

used by the Australian Academy of Science as part of *Australian Science, Australia's Future: Science 2035*, an initiative assessing the capability of the national science system and its role in achieving Australia's ambitions (Australian Academy of Science 2025).

The economic implications of workforce age structure shifts are well-documented (e.g., Bloom et al. 2007), affecting productivity, pensions and superannuation, and skill shortages (Productivity Commission 2013; OECD 2019a,b; Hyndman, Zeng & Shang 2021). The social implications are also significant, with an aging workforce leading to changes in workplace dynamics, potential problems with intergenerational knowledge transfer, and the need for policies that support older workers. Yet this problem does not appear to have been previously addressed from a statistical modelling perspective.

Our approach builds on functional data models, introduced to demographic modelling by Hyndman & Ullah (2007). They combined nonparametric smoothing and functional principal components for age-specific demographic rates. These models were then used by Hyndman & Booth (2008) for mortality, fertility, and migration rates, providing stochastic data generating processes for the components of demographic balance equations. These separate component models were then simulated to form future sample paths, leading to age- and sex-specific stochastic population forecasts. The modelling framework was later extended by Hyndman, Booth & Yasmeen (2013) to ensure coherence of forecasts between sexes or other demographic groups.

We propose a related approach for modelling workforce dynamics by redefining the demographic components in two ways. First, we replace fertility with workforce entry, which functions more like a migration process than a birth process because graduates can enter the workforce at any age. Second, we *explicitly* model workers leaving the workforce through two processes: retirement and death. Of course, people may also leave the workforce for other reasons, such as a career change or family commitments, but since we do not have data on these processes, we model them *implicitly* via a remainder term.

We describe the methodology in Section 2. By way of illustration, we apply the methodology to major scientific disciplines in Australia, focusing on the Natural and

Physical Sciences. We describe the data sources in Section 3, with the results provided in Section 4. The aim of this analysis is to inform future workforce planning and policy decisions to support the growth of Australia's scientific community. Finally, we provide some discussion and conclusions in Section 5.

2. Methodology

Suppose our workforce is divided into I groups, indexed by $i = 1, \dots, I$. In our application, these are scientific disciplines, but in principle they could refer to any subdivision of workers. Let $P_{i,x,t}$ denote the number of equivalent full-time workers in group i who are aged x at the start of year t, where $x = 15, 16, \dots$. The starting age of 15 is chosen because it is the minimum age at which individuals are counted as part of the labour force in the Australian Census (Australian Bureau of Statistics 2021b). We assume that data are available for years $t = 1, \dots, T$, and that forecasts are required for $P_{i,x,T+h}$ across all ages and groups, for some forecast horizon $h > 0$.

People can leave the workforce of a group through death, retirement, emigration, family responsibilities, or career change; they can enter the workforce through graduation, immigration, changes in family responsibilities, or career change. Unfortunately, we typically do not have data on many of these processes, so we will combine changes due to family responsibilities, career changes, emigration and immigration into a remainder term, which we denote as $E_{i,x,t}$. Let $D_{i,x,t}$ denote the number of deaths of workers in group i of age x in year t, $R_{i,x,t}$ denote the number of retirements from the same group of workers, and $G_{i,x,t}$ denote the number of new graduates of age x in year t who take up work in group i. The numbers in each case are for people aged x at the *start* of year t. Then population changes can be described using the following model:

$$P_{i,x+1,t+1} = P_{i,x,t} - D_{i,x,t} - R_{i,x,t} + G_{i,x,t} + E_{i,x,t}, \tag{1}$$

where

- $D_{i,x,t} \sim \text{Binomial}(P_{i,x,t}, q_{i,x,t})$, with $q_{i,x,t}$ being the probability of death for group i at age x in year t; and

- $R_{i,x,t} \sim \text{Binomial}((P_{i,x,t} - D_{i,x,t}), r_{i,x,t})$, with $r_{i,x,t}$ being the probability of retirement from group i at age x in year t.

That is, the population each year is equal to the population from the previous year having aged 1 year, minus the deaths or retirements that occurred during the previous year, plus the new graduates, plus any other changes due to migration or career change (which may be negative). We assume that $E_{i,x,t} = G_{i,x,t} = 0$ above some age threshold (say $x = 100$). Once $P_{i,x,t} = 0$ when x is above that threshold, all future populations $P_{i,x+k,t+k} = 0$, for $k = 1, 2, \dots$. That is, when the cohort aged x in year t has all retired or died, and x is above the threshold, they will not be replaced by new workers of the same age.

While our model was inspired by the stochastic population model of Hyndman & Booth (2008), that model has different inputs (births and immigration) and fewer outputs (deaths and emigration). Labour market forecasting is more complicated with no birth process, several more inputs (graduates, immigration, career changes, career renewal), and several more outputs (deaths, retirements, emigration, career disruption and career changes).

As a first approximation, the components q, r, E and G can be assumed to behave independently for each combination of i, x and t. In reality, there may be some negative correlation between G and E as insufficient graduates would probably lead to employers finding people from overseas, while too many graduates would lead to scientists seeking work elsewhere.

The choice of a Binomial rather than a Poisson distribution (in contrast to Brillinger (1986)) for deaths and retirements is because the Binomial distribution ensures that the number of deaths and retirements cannot exceed the population at risk. In a simulation context, with very small populations, this is important to avoid nonsensical results.

It is unlikely that we have available separate death and retirement counts for each group, and retirement data is not available in all years. So we will let $q_{i,x,t} = q_{x,t}$ and $r_{i,x,t} = r_x$, assuming that death rates and retirement rates are the same across all groups, and that retirement rates do not change over time. Similarly, graduation numbers are rarely available by discipline and age, so we will approximate $G_{i,x,t} = g_x G_{i,t}$ where $G_{i,t}$ is the total number of graduates in year t and g_x is the proportion of graduates by age across all disciplines.

This leads to the simpler model

$$P_{i,x+1,t+1} = P_{i,x,t} - D_{i,x,t} - R_{i,x,t} + g_x G_{i,t} + E_{i,x,t}, \tag{2}$$

where

- $D_{i,x,t} \sim \text{Binomial}(P_{i,x,t}, q_{x,t})$; and
- $R_{i,x,t} \sim \text{Binomial}((P_{i,x,t} - D_{i,x,t}), r_x)$.

For the age-specific time-varying components $q_{x,t}$ and $E_{i,x,t}$, we will use functional data models (Hyndman & Ullah 2007). For $q_{x,t}$, this is of the form

$$\log(q_{x,t}) = \mu(x) + \sum_{k=1}^{K} \beta_{k,t}\phi_k(x) + \varepsilon_{x,t}, \tag{3}$$

where $\mu(x)$ is the mean function, $\phi_k(x)$ are functional principal components, $\beta_{k,t}$ are the principal component scores, and $\varepsilon_{x,t}$ is an error term. Each $\beta_{k,t}$ is then modelled using a univariate time series model, such as an ARIMA model. A log-link function is used to ensure that the probabilities remain positive. An inverse logit link function could also be used, if the probabilities are close to 1 for some ages. A similar model is used for $E_{i,x,t}$, with separate mean functions and principal components for each group i. The number of components $K = 6$, for the reasons outlined in Hyndman & Booth (2008).

Because the $\beta_{k,t}$ values are principal component scores, they are uncorrelated by construction. While it is possible for there to be some cross-correlations between the series at lags other than zero, these are usually not large enough for a multivariate model to give more accurate forecasts (see, Hyndman & Ullah 2007) except in contrived simulated examples. On the other hand, Aue, Norinho & Hörmann (2015) did use multivariate models to capture these cross-correlations, although they did not compare them on real data.

Functional data models have been widely used in demography and other fields, and have been shown to work particularly well for age-specific demographic processes (Hyndman & Booth 2008; Booth et al. 2006). They enable the inherent smoothness over age to be captured, while modelling the autocorrelation over time using relatively simple univariate time series models applied to the principal component scores. In our application, we use univariate ARIMA models for the $q_{x,t}$ scores, and ARMA models for the $E_{i,x,t}$ scores, each estimated using maximum likelihood estimation. The assumption of stationarity for the $E_{i,x,t}$ scores is validated for the disciplines we consider, but is not a requirement in general.

For the time-varying component, $G_{i,t}$, we use a global ARIMA model (Hyndman & Montero-Manso 2021) to capture the dynamics over time and across disciplines. This is estimated using least squares estimation. The global model pools information across disciplines to improve forecast accuracy, especially for disciplines with limited historical data.

To forecast future working population numbers, $P_{i,x,t}$, $t > T$, we simulate future sample paths of each of the components $G_{i,t}$, $q_{x,t}$, and $E_{i,x,t}$, simulate $D_{i,x,t}$ and $R_{i,x,t}$ from their respective Binomial distributions, and then use the demographic growth-balance equation Equation 2 iteratively to obtain $P_{i,x,t}$ for $t = T+1, T+2, \dots$. This simulation-based approach allows us to capture the uncertainty in each of the components, leading to a distribution of possible future outcomes for $P_{i,x,t}$.

This model is somewhat pragmatic given the data available in our specific application. If better data were available, other variations on Equation 1 could be used. For example, if death rates were available by discipline, then we would replace $q_{x,t}$ by $q_{i,x,t}$ in the Binomial deaths distribution. If retirement rates were available by year, or by discipline, we could similarly replace r_x by a more specific retirement rate in the Binomial retirements distribution. If we had data on graduations by age and discipline, we could replace $g_x G_{i,t}$ by $G_{i,x,t}$. If we had data on migration, we could split the remainder $E_{i,x,t}$ into several components, and model them separately. None of this changes the overall modelling framework we are proposing.

Fortunately, there is no reason to think scientists of different disciplines would have different mortality experiences. A century ago, the dangers of radiation did increase mortality rates amongst chemists and physicists compared to other sciences, but modern science is conducted in extremely safe environments, so it seems reasonable to assume that all science disciplines share similar mortality profiles.

There has been a small increase in average retirement age over the last ten years due to an increase in the age at which the old age pension can be accessed (Hyndman, Zeng & Shang 2021), and a steady increase in the preservation age at which superannuation can be accessed (Kingston & Thorp 2019). However, there is no existing policy proposal to change either of these in the future, so it is reasonable to take the retirement age distribution in recent years as valid for the foreseeable future. Further, we know of no evidence that the socio-economic status of scientists varies with discipline, so there is no reason to think retirement intentions would change with discipline either.

We assume the age distribution of graduates is a product of age-dependent and time-dependent variables, g_x and $G_{i,t}$. Primarily, this is a pragmatic choice because we do not have more detailed data available. We can get age distributions of graduates across all disciplines in Australia, but not for each discipline; and we can get the numbers of graduates by discipline and year in Australia, but with no age breakdown. The most likely consequence of this simplifying assumption is that the variability in graduate numbers by age and time could be underestimated. It is conceivable that older graduates are drawn to different disciplines than younger graduates, or that fashionable disciplines change over time, resulting in different age distributions of the graduates over time. But without specific data related to this issue, we can only speculate.

It is also worth pointing out that the remainder term $E_{i,x,t}$ will absorb any inaccuracies that result from simplifying model assumptions in the other components, and we forecast the remainder allowing for changes over time, age and discipline. In fact, we could ignore all the model components and just forecast $P_{i,x,t}$ directly using a functional time series model, but that would fail to separate out the competing dynamics at play, and lead to much wider prediction intervals. By trying to model the individual components where we have available data, even if imperfectly, we capture more of the inherent uncertainty and obtain narrower prediction intervals.

3. Data

To illustrate the methodology, we consider the Natural and Physical Sciences as defined in the Australian Standard Classification of Education (ASCED) by the Australian Bureau of Statistics (2001). We refer to ASCED's Narrow Fields as "disciplines"; these comprise Physics and Astronomy, Mathematical Sciences, Chemical Sciences, Earth Sciences, Biological Sciences, Other Natural and Physical Sciences, and Natural and Physical Sciences not further defined (n.f.d.). Table 1 lists the detailed fields within each scientific discipline.

Table 1. Classification of scientific disciplines, based on the ASCED Narrow Fields of Education within the Broad Field of Natural and Physical Sciences. The table lists their corresponding Detailed Fields. "n.e.c." stands for "Not Elsewhere Classified."

Narrow Fields	**Detailed Fields**
Physics and Astronomy	Physics, Astronomy.
Mathematical Sciences	Mathematics, Statistics, Mathematical Sciences, n.e.c.
Chemical Sciences	Organic Chemistry, Inorganic Chemistry, Chemical Sciences, n.e.c.
Earth Sciences	Atmospheric Sciences, Geology, Geophysics, Geochemistry, Soil Science, Hydrology, Oceanography, Earth Sciences, n.e.c.
Biological Sciences	Biochemistry and Cell Biology, Botany, Ecology and Evolution, Marine Science, Genetics, Microbiology, Human Biology, Zoology, Biological Sciences, n.e.c.
Other Natural and Physical Sciences	Medical Science, Forensic Science, Food Science and Biotechnology, Pharmacology, Laboratory Technology, Natural and Physical Sciences, n.e.c.

We define the population of workers in a discipline as those who are active in the labour market and hold a bachelor's degree or higher in that discipline. For the purposes of this analysis, we will omit "Other Natural and Physical Sciences" and "Natural and Physical Sciences n.f.d.".

3.1. Working population

Data on the working population were sourced from the *Census of Population and Housing* (Australian Bureau of Statistics 2023) for census years 2006, 2011, 2016, and

2021. This dataset encompasses one-year age groups, the highest level of completed non-school qualification level (QALLP), the corresponding field of study (QALFP, Australian Bureau of Statistics 2021c), and the industries in which individuals work. However, labour force participation status (Australian Bureau of Statistics 2021a) is available only for 2016 and 2021. To estimate worker numbers for 2006 and 2011, the average participation rates from 2016 and 2021 were applied, assuming overall age distributions remain consistent.

The resulting estimates of the number of scientists who are active in the Australian labour market is shown in Figure 1 as the thick lines. Cohort interpolation (Stupp 1988), applying linear interpolation within each age cohort between census years, is used to estimate values for the intercensal years (shown as thin lines), giving $P_{i,x,t}$ for each discipline i, age x, and year t.

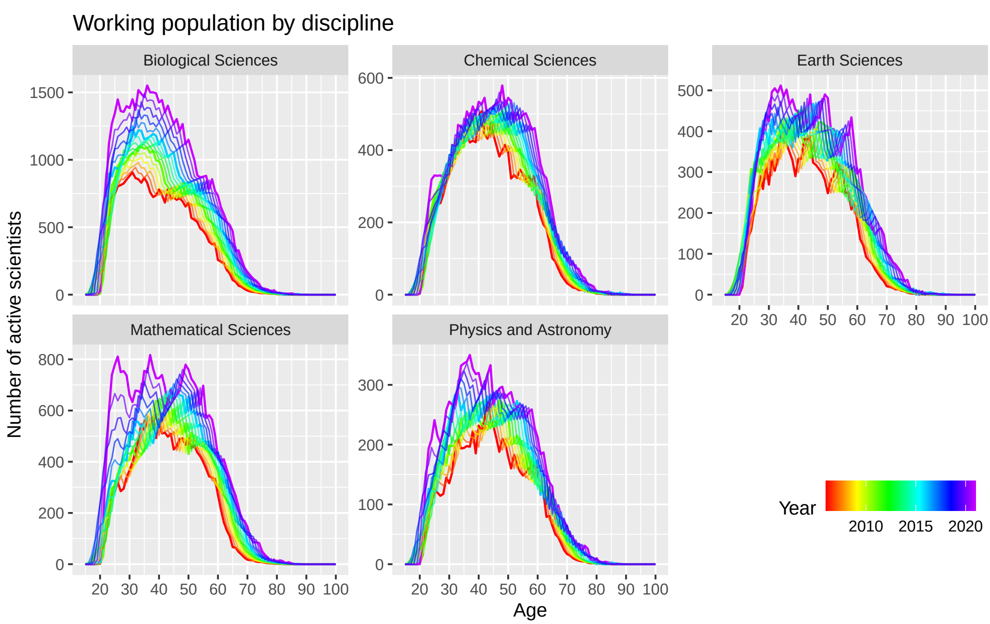

Figure 1. $P_{i,x,t}$: Estimated number of working scientists in Australia by discipline and age, 2006–2021. Thicker lines are used to denote census years.

3.2. Retirements

Retirement data was sourced from the *Retirement and Retirement Intentions* dataset (Catalogue 6238) for the 2022–2023 financial year (Australian Bureau of Statistics

2024). The data are categorised by the industry of an individual's main job, and are provided in four broad age groups (45–59, 60–64, 65–69 and 70+). There are 19 industry categories, with the largest numbers of scientists working in Education and Training (15.8%), Professional, Scientific and Technical Services (15.5%), and Health Care and Social Assistance (14.6%). The proportions in other industries are much smaller. We take a weighted average of retirement intentions using these top three industries, with proportions rescaled to sum to 1. The resulting values are shown in Figure 2 as the gray line. To obtain a single-year-of-age retirement distribution, we disaggregate the data using a monotonic cubic spline applied to the cumulative values of these age groups (Smith, Hyndman & Wood 2004). The resulting smoothed distribution (r_x) is shown as the black line in Figure 2.

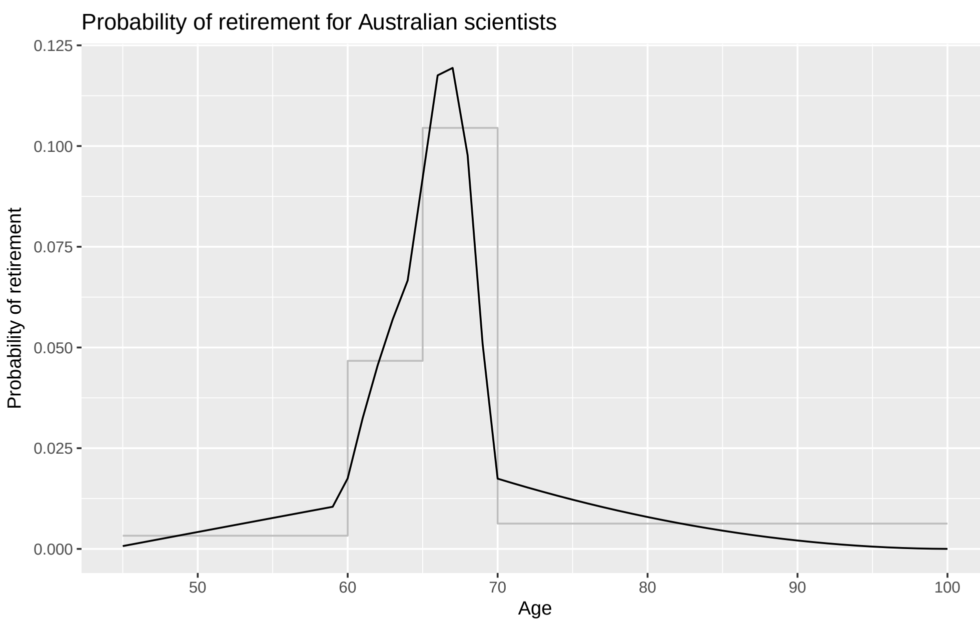

Figure 2. r_x: Age distribution of retirement intentions, based on data from the 2022–2023 Australian financial year. The grey line shows the age-group probabilities; the black line shows the smoothed probabilities.

3.3. Deaths

Age-specific mortality rates from 1971 to 2021 were obtained from the Human Mortality Database (2024). Using standard life table methods, these rates are converted into age-specific probabilities of death, as shown in Figure 3. Over time, mortality probabilities have generally declined across all age groups, reflecting improvements in Australian life expectancy.

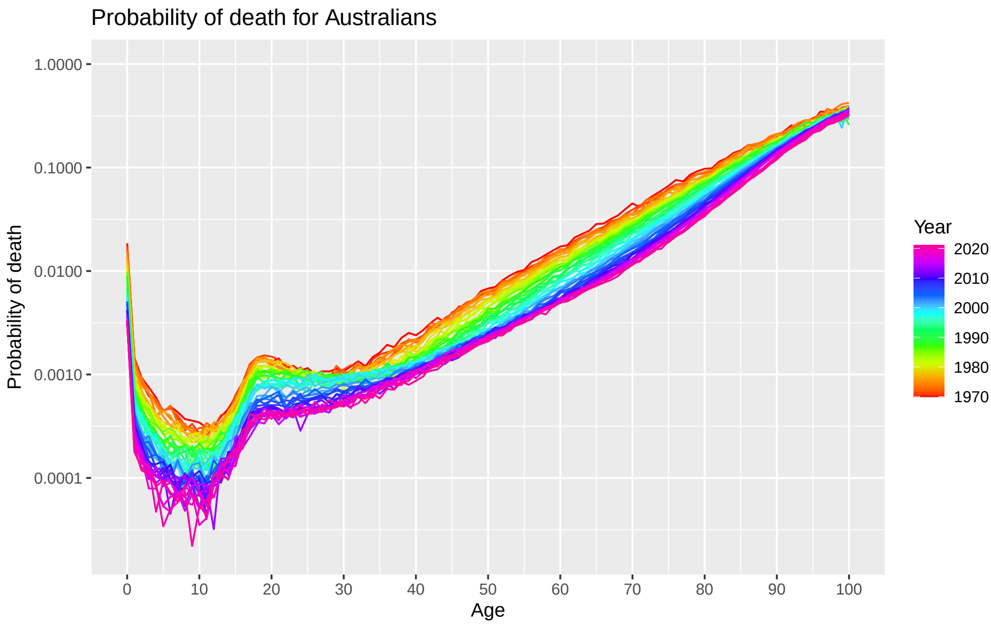

Figure 3. $q_{x,t}$: Age-specific probabilities of death (on a logarithmic scale) for each year from 1971 to 2021.

No data are available for specific industry groups, so we assume that all scientists have the same mortality probabilities as the general population. These probabilities serve as estimates of $q_{x,t}$.

3.4. Graduate completions

Graduate completion statistics were obtained from the *Award Course Completions* dataset (Department of Education 2024b). Figure 4 shows the distribution of graduate completions with a bachelor's degree or higher, by age for each year from 2006 to 2023. Some missing values result in gaps in certain lines, but the overall pattern remains highly consistent across years. Given this consistency, the data is averaged across all available years, and then smoothed by applying monotonic cubic splines to the cumulative values (Smith, Hyndman & Wood 2004). The resulting averaged distribution, shown as the black line in Figure 4, smooths out year-to-year fluctuations and provides an estimate of g_x.

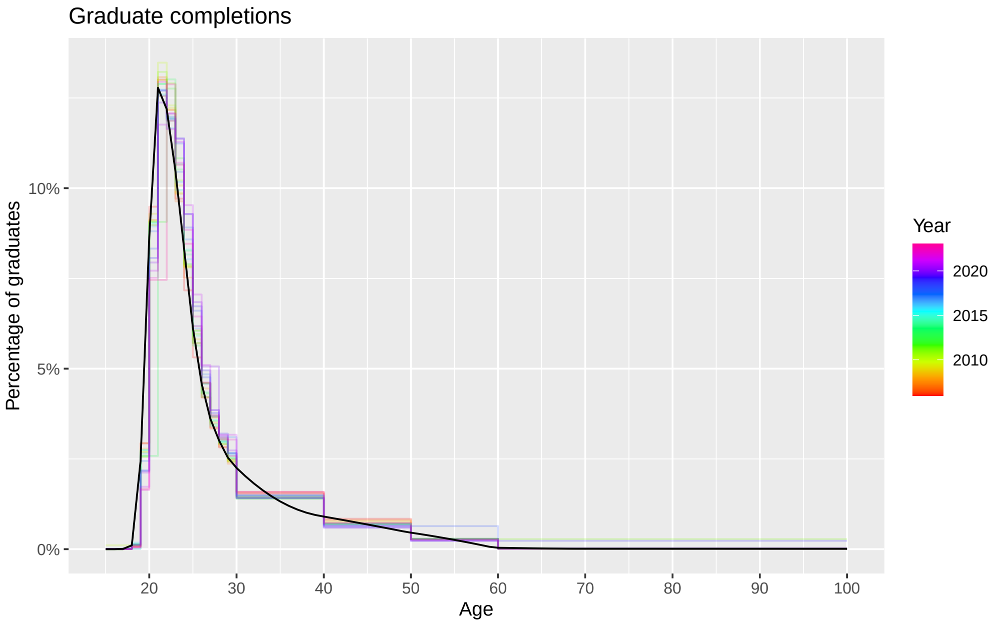

Figure 4. g_x: Estimated distribution of graduate completions by age (black). This is estimated by averaging and smoothing the data for the years 2006 to 2023 (coloured).

The Department of Education provides data on the number of graduates with a bachelor's degree or higher, categorised by discipline and year (Department of Education 2024a). This dataset includes both domestic and international students. The total number of graduates, $G_{i,t}$, in each discipline i and year t, are shown in Figure 5.

The large increase in the working population observed in the 2021 Census for Mathematical Sciences (Figure 1) can be partly attributed to the sharp rise in graduate numbers between 2016 and 2021. This is probably due to the impact of data science, and the growing importance of statistics and machine learning in many areas of employment.

3.5. Remainder

The demographic growth-balance equation (Equation 2), when rearranged, provides an expression for the remainder including net migration and career changes:

$$E_{i,x,t} = P_{i,x+1,t+1} - P_{i,x,t} - D_{i,x,t} - R_{i,x,t} - g_x G_{i,t}, \tag{4}$$

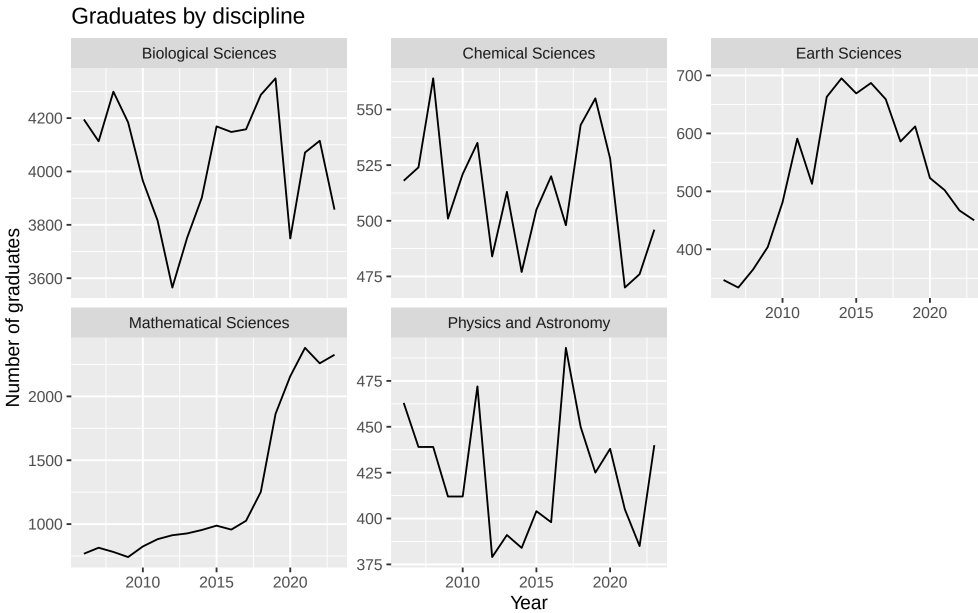

Figure 5. $G_{i,t}$: Total number of graduates with a bachelor's degree or higher by discipline from 2006 to 2023.

However, we do not have data on $D_{i,x,t}$ and $R_{i,x,t}$, so we replace these by their expected values, $P_{i,x,t}q_{x,t}$ and $P_{i,x,t}(1-q_{x,t})r_x$, respectively. We can only estimate remainders up to 2020 because we need data for both year t and year $t+1$ in Equation 4, and our working population data only extends to 2021. The estimated remainders are shown in Figure 6.

The inclusion of international students in the graduate data leads to large positive values of the remainder for the teenage years, followed by large negative values when these students return to their home countries after graduation.

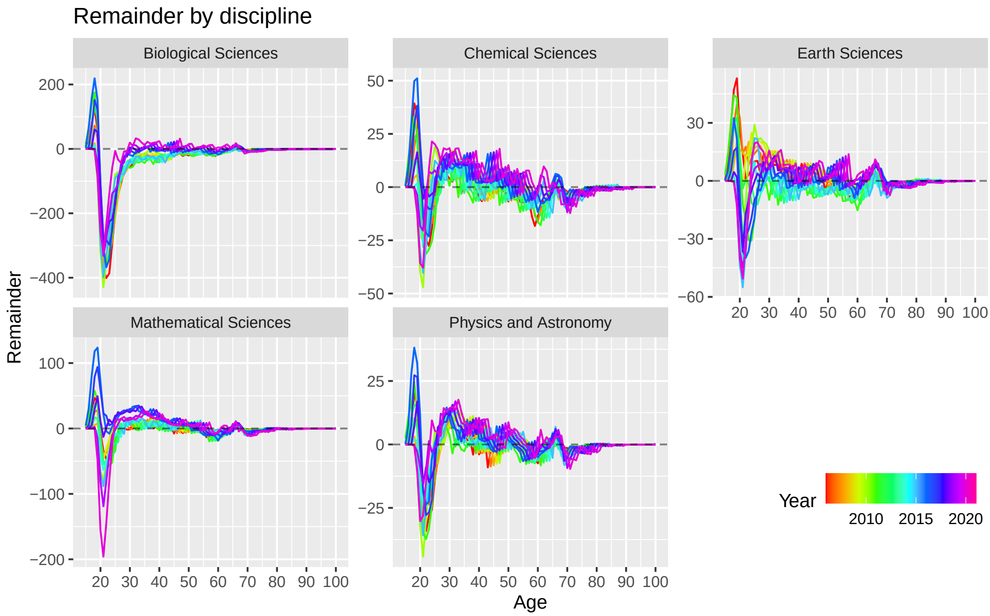

Figure 6. Estimated remainder $E_{i,x,t}$ by discipline, age and year (2006–2020).

4. Results

4.1. Graduate completions

To forecast future graduate numbers, $G_{i,t}$, a global ARIMA model was employed, following the principles outlined by Hyndman & Montero-Manso (2021). The global model captures overall trends across disciplines by scaling graduate data within each discipline, ensuring proportional contributions from all disciplines before fitting the global ARIMA model. This improves the numerical stability of the model by incorporating information across disciplines. The forecast distributions are shown in Figure 7, with the mean forecast represented by the solid line and 90% prediction intervals indicated by the shaded area.

4.2. Death probabilities

The death probabilities shown in Figure 3 were first smoothed using the partially monotonic penalised spline approach of Hyndman & Ullah (2007). Then the functional data model Equation 3 was estimated, with ARIMA models fitted to the coefficients. The forecasts for one year are shown in Figure 8, with the mean forecast represented

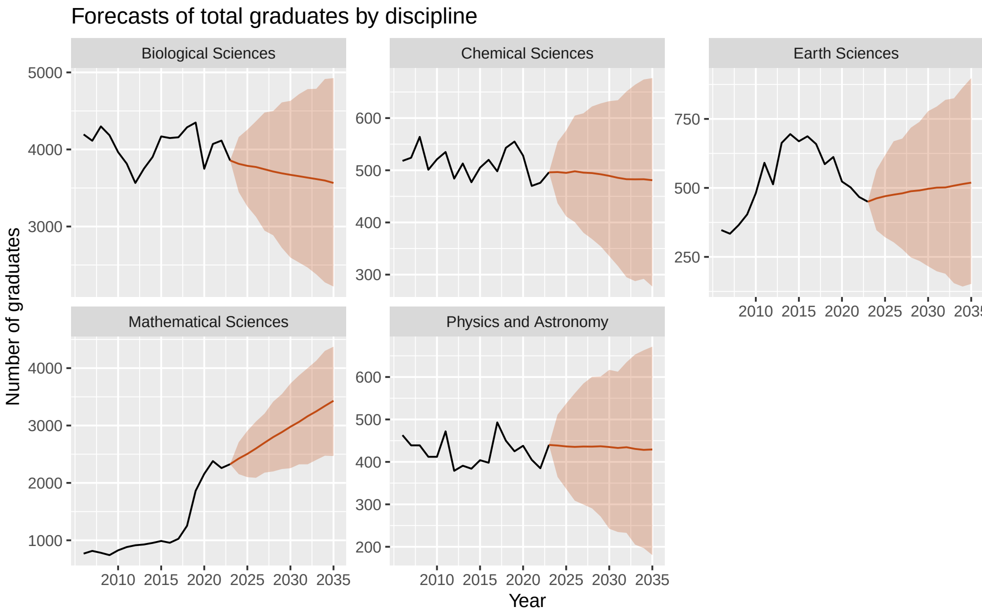

Figure 7. Forecast of $G_{i,t}$: the number of graduates by discipline, 2024–2035, based on historical data from 2006–2023. The shaded regions represent the 90% prediction intervals, and the solid lines indicate the mean estimates.

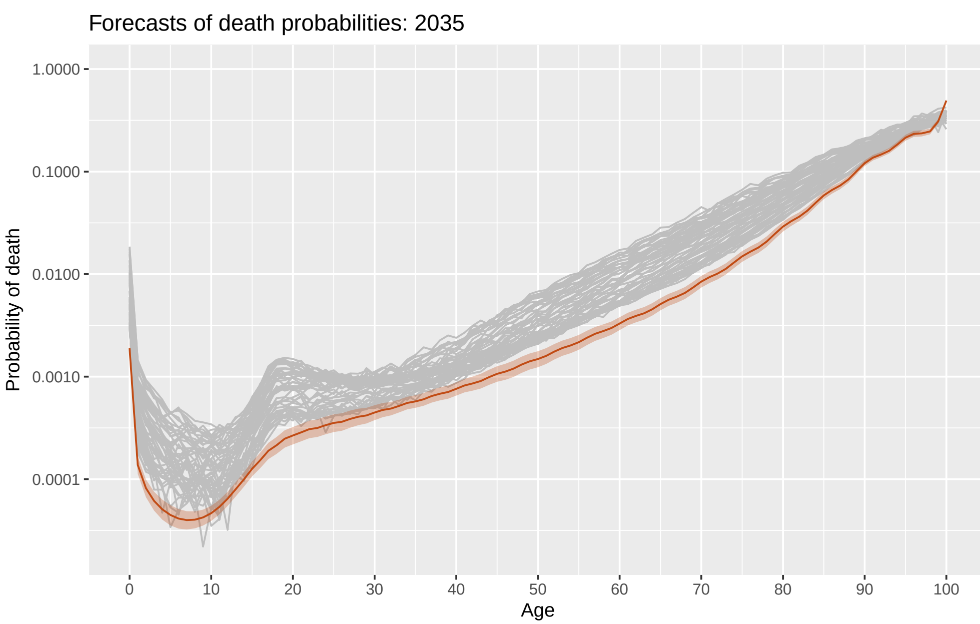

Figure 8. Forecasts of $q_{x,t}$: age-specific probabilities of death (on a logarithmic scale) for 2035, based on historical data from 1971–2021. The shaded regions represent the 90% prediction intervals, and the solid lines indicate the mean estimate.

by the solid line and 90% prediction intervals indicated by the shaded area. Note that the historical data (shown in gray) represent unsmoothed values, while the forecasts are based on the smoothed functional data model. The additional variation seen in the historical data is captured in the model through the Binomial death process.

4.3. Remainder

The remainder, $E_{i,x,t}$, is also modelled using a functional data model (Hyndman & Ullah 2007), with ARIMA models fitted to the principal component scores. In this case, all scores were found to be stationary using the KPSS test (Kwiatkowski et al. 1992), so ARMA models are used. The forecasts for one year are shown in Figure 9, with the mean forecast represented by the solid line and 90% prediction intervals indicated by the shaded area.

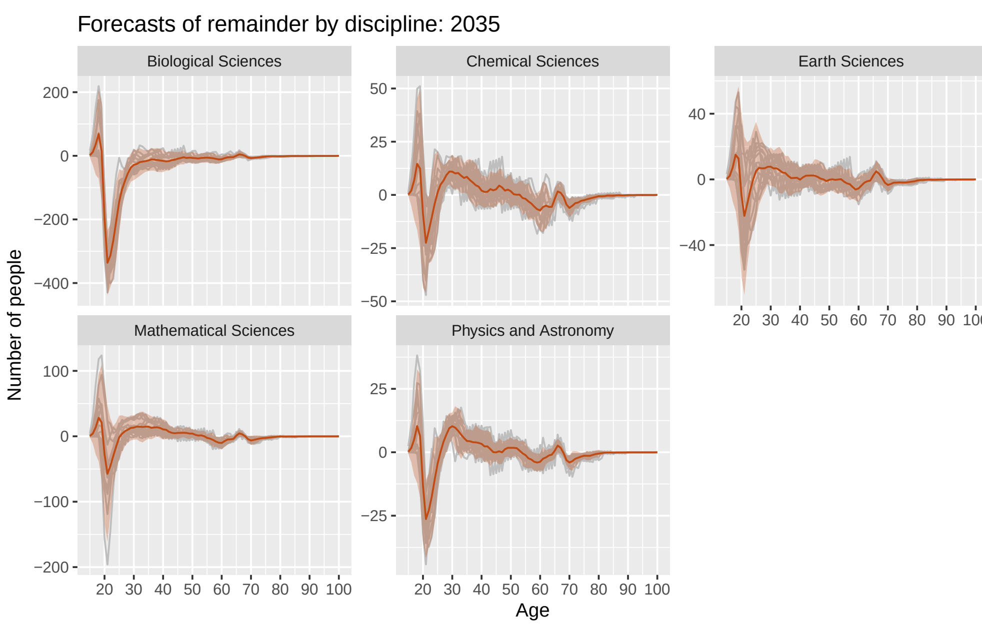

Figure 9. Forecasts of $E_{i,x,t}$: the remainder by discipline for 2035, based on historical data from 2006–2020. The shaded regions represent the 90% prediction intervals, and the solid lines indicate the mean estimates.

4.4. Simulating future populations

We use the demographic growth-balance model Equation 2 to iteratively simulate future populations, using the models described above for the components. The following steps outline the process.

A total of 1000 simulations are run to obtain a distribution of future age-specific population scenarios. The average of the 1000 simulations provides the mean age-specific forecast, while quantiles estimate forecast uncertainty. Figure 10 presents the mean and 90% prediction intervals for 2035.

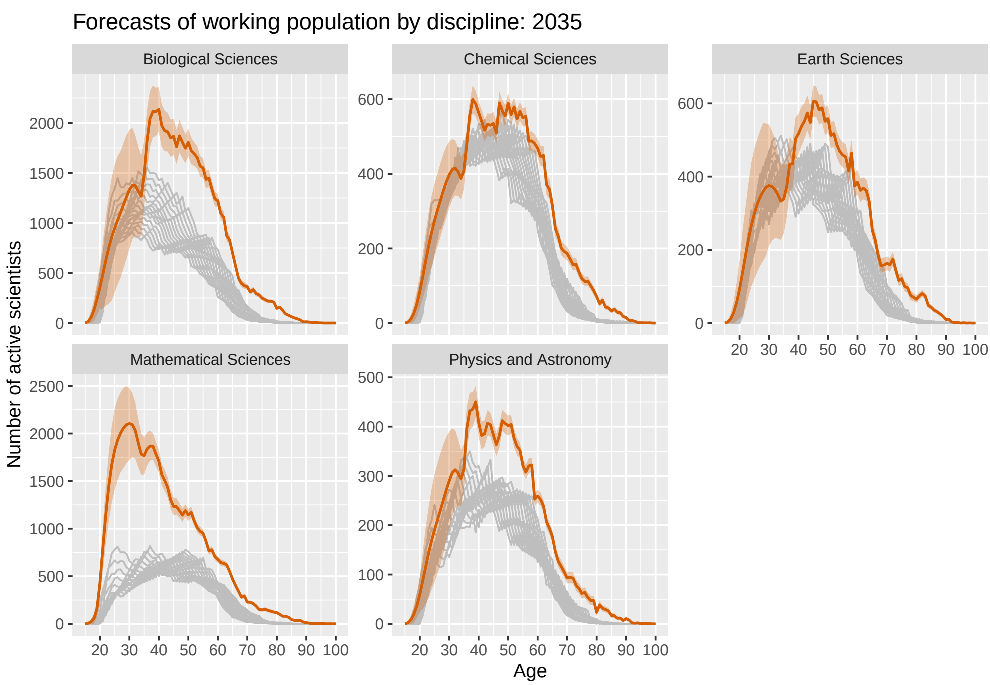

Figure 10. Forecasts of $P_{i,x,t}$: the working population by discipline for 2035. The shaded regions represent the 90% prediction intervals, and the solid lines indicate the mean estimates.

In 2035, forecast variability is highest in the age period 20–35 years, before gradually narrowing as the workforce ages. This is due to the relatively high uncertainty in the new graduates component compared to the other components. Mid-to-late career estimates primarily reflect the aging of existing cohorts. The prediction intervals become especially narrow during the retirement phase, where the workforce dynamics become more predictable. Since retirements increase after the late 50s, workforce participation beyond 60 serves as a benchmark for identifying trends in delayed retirement and extended career duration. Over the next ten years, we expect an aging workforce in all but the Mathematical Sciences, where a large increase in the population is forecast.

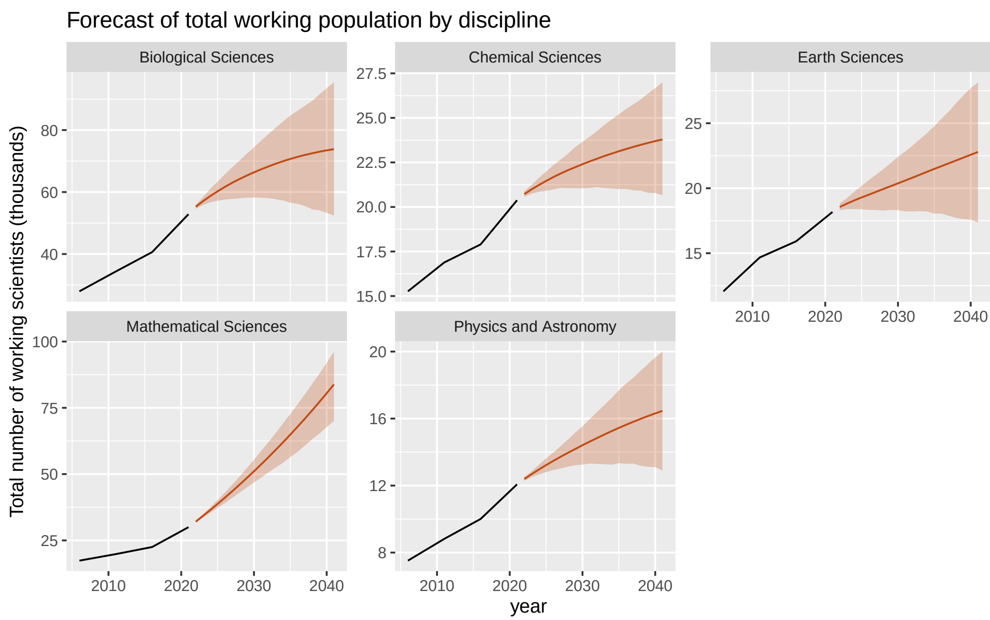

Figure 11. Forecasted total number of working scientists across scientific disciplines from 2022 to 2041. The shaded region represents the 90% prediction interval, the coloured line indicates the mean estimate, and the black line represents historical data.

Cohort effects are also visible in Figure 10, where fluctuations in earlier ages and years propagate through to later ages and years. This is particularly evident in the mid-career years because there are few deaths and retirements, few graduates older than 30, and the variation due to the remainder term is relatively small after age 30.

Summing over ages allows for estimating the forecast distribution of the total number of working scientists in each future year, as shown in Figure 11. The forecasts indicate continued growth, but at a gradually slower pace for all disciplines other than the Mathematical Sciences. Where the lower bound is nearly flat, workforce stagnation is possible in a conservative scenario. Even in the optimistic scenario (corresponding to the upper bound), growth only slightly exceeds the current pace, except for the Mathematical Sciences. As noted earlier, the divergent behaviour of the Mathematical Sciences is likely driven by the growing importance of data science and related fields.

The wide prediction intervals reflect the uncertainty in the forecasts, and show that caution is needed when interpreting the results. The only discipline where there is clear evidence of growth or decline is the Mathematical Sciences. For all other disciplines, the prediction intervals include the current level, indicating that stagnation, increase, or decline is possible.

5. Discussion

While these forecasts provide a foundation for workforce planning, it is important to note that they are entirely driven by historical trends and do not account for possible new developments, such as the impact of AI and other emerging technologies on the labour market in different scientific disciplines. Other factors, such as policy changes or global economic shifts, may also influence workforce trends. These exogenous factors could be accounted for by adding covariates into the time series models used, provided relevant data are available.

Evaluating and validating these forecasts is challenging due to the relatively long forecast horizon compared to the available historical data. While time series cross-validation (Hyndman & Athanasopoulos 2021) could be used to assess forecast accuracy for shorter horizons, the benefit of the forecasts is primarily for longer horizons, where such validation is not possible. The uncertainty in the forecasts, as reflected in the wide prediction intervals, highlights the need for caution when interpreting the results.

If more detailed data were available, the model could be refined further by including, for example, discipline-specific death rates, retirement data by year and/or discipline, graduate data by age and discipline, and data on migration and career changes. It is not clear how much these refinements would improve forecast accuracy, but they would likely reduce uncertainty in the forecasts.

Forecasts are often designed not just to predict the future, but also to inform policy decisions, and so modify the future. In this context, these forecasts could be used to identify potential skill shortages or surpluses in specific disciplines, guiding decisions on university and immigration policy, and thus changing the future outcomes (Hyndman

2023). Consequently, forecast accuracy may be less important than understanding the range of possible outcomes and their implications for policy.

While this analysis has focused on the scientific workforce in Australia, the methodology could be applied to other countries or workforce sectors, provided similar data are available. The specific components of the demographic growth-balance equation may need to be adapted to reflect the available data in other applications.

6. Software and reproducibility

All results presented here can be reproduced using the code available at `https://github.com/robjhyndman/age_structure_forecasts`. The analysis was conducted using R version 4.5.1 (R Core Team 2025), with the following R packages: vital (Hyndman et al. 2025), tsibble (Wang et al. 2025), fable (O'Hara-Wild et al. 2024), targets (Landau 2025, 2021), ggplot2 (Wickham et al. 2025; Wickham 2016), and other tidyverse (Wickham et al. 2019) packages.

References

Aue, A., Norinho, D.D. & Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association* **110**, 378–392.

Australian Academy of Science (2025). Australian science, australia's future: Science 2035. URL `https://www.science.org.au/supporting-science/australian-science-australias-future-science-2035`.

Australian Bureau of Statistics (2001). Broad, narrow and detailed fields. URL `https://www.abs.gov.au/statistics/classifications/australian-standard-classification-education-asced/2001/field-education-structure-and-definitions/structure/broad-narrow-and-detailed-fields`.

Australian Bureau of Statistics (2021a). Labour force participation flag (LFFP). URL `https://www.abs.gov.au/census/guide-census-data/census-dictionary/2021/variables-topic/national-reporting-indicators/labour-force-participation-flag-lffp`.

Australian Bureau of Statistics (2021b). Labour force status (LFSP). URL `https://www.abs.gov.au/census/guide-census-data/census-dictionary/2021/variables-topic/income-and-work/labour-force-status-lfsp`.

Australian Bureau of Statistics (2021c). Non-school qualification: field of study (QALFP). URL `https://www.abs.gov.au/census/guide-census-data/census-dictionary/2021/variables-topic/education-and-training/non-school-qualification-field-study-qalfp`.

Australian Bureau of Statistics (2023). Microdata and tablebuilder: Census of population and housing. URL https://www.abs.gov.au/statistics/microdata-tablebuilder/available-microdata-tablebuilder/census-population-and-housing. (2006, 2011, 2016, 2021).

Australian Bureau of Statistics (2024). Retirement and retirement intentions, Australia. URL https://www.abs.gov.au/statistics/labour/employment-and-unemployment/retirement-and-retirement-intentions-australia/latest-release. Table 9.2.

Bloom, D.E., Canning, D., Fink, G. & Finlay, J.E. (2007). Does age structure forecast economic growth? *International Journal of Forecasting* **23**, 569–585.

Booth, H., Hyndman, R.J., Tickle, L. & de Jong, P. (2006). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research* **15**, 289–310.

Brillinger, D.R. (1986). The natural variability of vital rates and associated statistics. *Biometrics* **42**, 693–734.

Department of Education (2024a). Award course completions. URL https://www.education.gov.au/higher-education-statistics/higher-education-statistics-data. Higher Education Data Request (2006-2023).

Department of Education (2024b). Award course completions for all students by age group and broad level of course. URL https://www.education.gov.au/higher-education-statistics/student-data. Table 5 (2006-2016), Table 14.5 (2017-2023).

Human Mortality Database (2024). Human mortality database. URL https://www.mortality.org. Accessed on 2024-12-18.

Hyndman, R.J. (2023). Forecasting, causality and feedback. *International J Forecasting* **39**, 558–560.

Hyndman, R.J. & Athanasopoulos, G. (2021). *Forecasting: principles and practice*. Melbourne, Australia: OTexts, 3rd edn. URL https://www.OTexts.com/fpp3/.

Hyndman, R.J. & Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International J Forecasting* **24**, 323–342.

Hyndman, R.J., Booth, H. & Yasmeen, F. (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography* **50**, 261–283.

Hyndman, R.J. & Montero-Manso, P. (2021). Principles and algorithms for forecasting groups of time series: Locality and globality. *International J Forecasting* **37**, 1632–1653.

Hyndman, R.J., Tang, S., McBain, M. & O'Hara-Wild, M. (2025). *vital: Tidy Analysis Tools for Mortality, Fertility, Migration and Population Data*. URL https://pkg.robjhyndman.com/vital/.

Hyndman, R.J. & Ullah, S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis* **51**, 4942–4956.

Hyndman, R.J., Zeng, Y. & Shang, H.L. (2021). Forecasting the old-age dependency ratio to determine a sustainable pension age. *Australian & New Zealand J Statistics* **63**, 241–256.

Kingston, G. & Thorp, S. (2019). Superannuation in australia: A survey of the literature. *Economic Record* **95**, 141–160.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. how sure are we that economic time series have a unit root? *Journal of Econometrics* **54**, 159–178.

Landau, W.M. (2021). The targets R package: a dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software* **6**, 2959.

Landau, W.M. (2025). *targets: Dynamic Function-Oriented 'Make'-Like Declarative Pipelines*. URL https://cran.r-project.org/package=targets.

OECD (2019a). OECD employment outlook 2019: The future of work. URL https://www.oecd.org/en/publications/oecd-employment-outlook-2019_9ee00155-en.html. Accessed on 2025-04-16.

OECD (2019b). Working better with age. URL https://www.oecd.org/en/publications/working-better-with-age_c4d4f66a-en.html. Accessed on 2025-04-16.

O'Hara-Wild, M., Hyndman, R.J., Wang, E., Caceres, G., Bergmeir, C., Hensel, T.G. & Hyndman, T. (2024). *fable: Forecasting Models for Tidy Time Series*. URL https://fable.tidyverts.org.

Productivity Commission (2013). An ageing Australia: preparing for the future. URL https://www.pc.gov.au/research/completed/ageing-australia/ageing-australia-overview.pdf. Accessed on 2025-04-16.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Smith, L., Hyndman, R.J. & Wood, S.N. (2004). Spline interpolation for demographic variables: the monotonicity problem. *J Population Research* **21**, 95–98.

Stupp, P.W. (1988). Estimating intercensal age schedules by intracohort interpolation. *Population Index* **54**, 209–224.

Wang, E., Cook, D., Hyndman, R.J., O'Hara-Wild, M., Smith, T. & Davis, W. (2025). *tsibble: Tidy Temporal Data Frames and Tools*. URL https://tsibble.tidyverts.org.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, USA: Springer-Verlag. URL https://ggplot2-book.org/.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686.

Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., van den Brand, T. & Posit, PBC (2025). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. URL https://cran.r-project.org/package=ggplot2.