Forecasting the age structure of the scientific workforce in Australia

We thank the reviewers for their helpful comments. We have included all reviewer comments in blue below, followed by our responses in black.

Reviewer 1

This manuscript presents a statistically rigorous and highly valuable approach for forecasting the age structure of specialized workforces. The proposed methodology is robust, leveraging functional data models for age-specific rates and employing a global ARIMA model to enhance graduate forecast accuracy. A simulation-based approach captures comprehensive uncertainty. The manuscript is written well, organized clearly. This manuscript is suitable for the *Australian & New Zealand Journal of Statistics* due to its innovative extension of established functional data models and its direct application to critical national policy issues.

Thanks for your positive feedback.

Major comments

1. Regarding the model assumptions, the main contribution of this study lies in a practical modification of the traditional population growth balance equation. As shown in equation (2), it is assumed that the mortality rate and retirement rate are uniform across all scientific disciplines (group i), and that the retirement rate is constant over time ($q_{i,x,t} = q_{x,t}$ and $r_{i,x,t} = r_x$). It is suggested that the potential implications of these assumptions be more explicitly discussed. For instance, are there no significant differences in the deaths or retirements in the group 'Chemical Sciences' compared to those in the group 'Mathematical Sciences'?

We assume that mortality rates change with age and time, but not with discipline; and that retirement rates change with age, but not with time and discipline. There are two reasons for these assumptions. First, there is the purely pragmatic reason that the data we have available does not allow anything else. Mortality rates by scientific discipline are not available anywhere in the world, to our knowledge, so the best we can do with the available data is to use the same mortality rates for all disciplines, while allowing for change over time and age. 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.

Retirement data is even more difficult to come by, and we have only been able to obtain retirement data by age for a single year in broad age groups. 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, and a steady increase in the preservation age at which superannuation can be accessed. 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.

The referee asks specifically about the Chemical Sciences and Mathematical Sciences groups. It is not clear to us why these have been singled out, but we know of no evidence that retirement or mortality rates would differ between these groups.

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.

We have added new material in Section 2 incorporating some of these comments.

1. The authors mentioned that the age distribution of graduates entering the labor market (g_x) was estimated in the paper by averaging and smoothing the data across all available years (2006 to 2023). Does the assumption in this paper that g_x is fixed and time-invariant ($g_{x,t} = g_x$) potentially underestimate long-term dynamic changes? It is suggested that the impact of this assumption on the variability of young population forecasts be discussed either in the Methodology or Results section.

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. In any case, the comment above regarding the remainder term is also relevant here.

We have added a new paragraph in Section 2 along these lines.

3. The authors mentioned that the remainder includes large positive and negative fluctuations caused by international students returning to their home countries after graduation, as well as complex economic behaviors involving career changes and international migration. Strictly speaking, are these macroeconomic and social trends ($E_{i,x,t}$) thought to be long-term stationary? It is recommended that the discussion take into account that the stationarity assumption will fail if major policy or global economic structural changes take place, and this is a significant factor in forecasting uncertainty.

We do not assume that the remainder term is stationary. We used stationary ARMA processes for the corresponding principal component scores because they all were found to be stationary using a KPSS test. But there is no reason why they could not be modelled using non-stationary processes if the data supported it. In particular, if major policy or global economic structural changes took place, this is likely to lead to non-stationarity in the remainder term, leading to greater forecast uncertainty.

We have clarified this point in Section 3.

Reviewer 2

In this paper, the authors studied the forecasting of the workforce of different scientific disciplines. The forecasting approach was built upon early studies by the authors with a reminder term to account for migration and career changes. Although the empirical study in Sections 3 and 4 may be interesting to applied scientists, the statistical contributions of the paper are unclear. More detailed comments are given below.

Other comments:

1. On page 2, it seems restrictive to have only two reasons, retirement and death, for people to leave the workforce.

We do not have only these two reasons for people to leave the workforce. We also mention career changes, migration, and career disruption. These are handled implicitly in the remainder term, as we do not have data available for these processes. Deaths and retirement can be handled explicitly as we have suitable data available.

We have added clarifying statements in Section 2 to address this misunderstanding.

2. On page 4, assuming $q_{i,x,t} = q_{x,t}$ and $r_{i,x,t} = r_{x,t}$ needs justification rather than for modeling convenience.

See the response to Reviewer 1, comment 1.

1. The main model setup is stated on pages 6 and 7. It is not clear which parts of the model are new in the literature.

To the best of our knowledge, there has been no previous attempt to forecast the age-structure of a workforce using a stochastic model. So the model is entirely new. We apologize if we gave a different impression in discussing the connection between our proposed model and the population forecasting model of Hyndman and Booth (IJF, 2008). Both involve age-structure forecasting, but with very different processes. We are proposing a new model for workforce age structure forecasting, that is somewhat inspired by the statistical model proposed by Hyndman and Booth (IJF, 2008) for population forecasting. The latter involves different inputs (births and immigration) and fewer outputs (deaths and emigration). Labour market forecasting is more complicated with no birth process, but several more inputs (graduates, immigration, career changes, career renewal) and several more outputs (deaths, retirements, emigration, career disruption and career changes).

We have now made this clearer in Section 2, and in the abstract.

4. In equation (3), why don't the authors consider a multivariate time series model for $\beta_{k,t}$, $k = 1, ..., K$?

The $\beta_{k,t}$ values are principal component scores, so 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 [1] except in contrived simulated examples. On the other hand, A. Aue, D. D. Norinho, and S. Hörmann [2] did use multivariate models to capture these cross-correlations, although they did not compare them on real data.

We have added a new paragraph along these lines in Section 2.

5. There is no mention of the statistical inference of the unknown parameters. For example, how were the unknown parameters in the model on pages 6 and 7 estimated?

Functional principal components [1] was used to compute the principal component scores. All univariate ARIMA models were estimated using maximum likelihood estimation. The global ARIMA model was estimated using least squares estimation (as a full likelihood is somewhat difficult to compute across multiple series).

This has now been noted in Section 2.

6. There is no simulation experiment to investigate the performance of the inference proce- dure.

Statistical inference is not the point of this paper. We estimate model parameters in order to produce forecasts. The values of the parameters themselves are of little interest. In any case, the functional time series models used for each component have been well-studied previously [1], [2], [3], [4], [5].

7. In Section 4, the forecasting was done using simulation methods. There is no validation of the forecasting accuracy. It is important to demonstrate how good the forecasts are compared with existing models.

Which existing models? There are no existing statistical models for forecasting labour market age-structure as far as we know, so we can't compare against them.

However, we have added a new appendix validating the coverage of the one-step prediction intervals for the last year of census data, using data up to the previous census. Given the use of interpolation between censuses, it would be invalid to test the accuracy of our forecasts between censuses, and there is insufficient data to properly evaluate forecasts earlier than the last census.

In short, the contributions of the current version of the paper do not seem to be adequate for publication in the journal. Much more work is required to illustrate the value of the proposed changes in the workforce modeling.

It is not clear what is meant by "proposed changes in the workforce modeling". We are not proposing changes, we are proposing an entirely new model — the first statistical model of any kind to allow age structure forecasting for a labour market. Hopefully the comments above, and the new material in the paper, will have made this clearer.

Bibliography

[1] R. J. Hyndman and S. Ullah, "Robust forecasting of mortality and fertility rates: A functional data approach," *Computational Statistics & Data Analysis*, vol. 51, no. 10, pp. 4942–4956, 2007.

[2] A. Aue, D. D. Norinho, and S. Hörmann, "On the Prediction of Stationary Functional Time Series," *Journal of the American Statistical Association*, vol. 110, no. 509, pp. 378–392, 2015, doi: 10.1080/01621459.2014.909317.

[3] R. J. Hyndman and H. L. Shang, "Forecasting functional time series (with discussion)," *J Korean Statistical Society*, vol. 38, no. 3, pp. 199–221, 2009.

[4] L. Horváth and P. Kokoszka, *Inference for functional data with applications*. New York: Springer, 2012.

[5] H. Haghbin and M. Maadooliat, "A journey from univariate to multivariate functional time series: A comprehensive review," *Wiley interdisciplinary reviews. Computational statistics*, 2023, doi: 10.1002/wics.1640.