Australian Government

National Skills Commission

Introducing NERO: A monthly Nowcast of Employment by Region and Occupation (NERO) in Australia

Samuel Shamiri, Leanne Ngai, Peter Lake,
Yin Shan, Amee McMillan, Therese Smith

This paper outlines the methodology used to develop a Nowcast of Employment by Region and Occupation (NERO) in Australia. Nowcasting is an emerging technique in the field of economics used mainly to derive data on economy-wide indicators such as GDP and unemployment. This paper shows how the NSC has applied nowcasting to the labour market at a detailed level, producing a new monthly series of employment by region and occupation in Australia. Complementing existing data sources, this new series will improve Australia's capacity to understand labour market trends in a more timely and detailed manner.

5 July 2021

Acknowledgements

Sincere appreciation must be extended to Bjorn Jarvis of the Australian Bureau of Statistics (ABS) and Bilal Rafi and Adam Bialowas from the Labour Economics Section of the Department of Education, Skills and Employment (DESE) for their support during this project.

Introduction

The impacts of COVID-19 on economies around the world have demonstrated the need for timely, detailed and accurate labour market data to support targeted monitoring and policy interventions. Existing data on occupational employment by region were not originally designed to support the frequent and detailed measurement becoming increasingly important in times of uncertainty. With this in mind, this paper details the methodology the National Skills Commission (NSC) used to develop a new Nowcast of Employment by Region and Occupation (NERO) dataset. Through the use of innovative techniques combining traditional economic and real-time data sources, the NSC's NERO provides a new data asset to help support more responsive decision-making.

1. What is nowcasting?

The NSC uses real time datasets, as well as big data techniques to deliver NERO on a monthly interval. Unlike forecasting, nowcasting does not attempt to predict or anticipate the future – its focus is understanding the now.

The goal of nowcasting is to produce a more frequent estimate of an economic series to support more responsive decision-making. This may be a timelier estimate of GDP (Higgins, 2014), the unemployment rate (Moriwaki, 2020) or current economic trends (OECD, 2017; Kindberg-Hanlon and Sokol, 2018; Bok et al, 2017; Nguyen and La Caga, 2020).

In economics, nowcasting has often been undertaken using time series econometric techniques and statistics, including vector autoregressions and mixed sampling methods. However, innovations on two fronts are transforming how nowcasting analysis is being undertaken. One relates to the availability of new, novel data sets. The other is the emergence of machine learning techniques in economic analysis (Varian, 2014). These two innovations are extending the reach of nowcasting into new fields, including labour market analysis. As noted by Dawson et al (2020) "the confluence of more available labour market data facilitated by the internet (for example job ads), advances in computation and greater access to analytical tools (such as machine learning) are enabling more data-driven approaches for the labour prediction tasks". This provides a new way of potentially examining labour market activity.

2. Limitations of existing employment data by occupation and region in Australia

Australian industries, occupations, and regions vary considerably. Understanding this variation in a timely manner through nowcasting, is the aim of NERO.

Robust estimates of employment by occupation and region are available from the ABS Census of Population and Housing, however, these data are only collected every five years and arrive with a relatively long time-lag after collection.

Data is available from the Household, Income and Labour Dynamics in Australia (HILDA) Survey, however, this survey is only conducted on an annual basis and the data are often very sparse at a detailed geographic level due to the relatively small sample size of the survey (approximately 17,000).[1] This data also arrives with a relatively long time-lag after its collection. Given it is a longitudinal survey, it is also benchmarked to the ABS Labour Force Survey, to minimise bias in its cross-sectional estimates.

The ABS provided the NSC with quarterly estimates of employment by region and occupation, from the Labour Force Survey to support the NSC's NERO project. While the Labour Force Survey has a large sample of around 50,000 people, around 30,000 of whom are employed, it also faces sparsity and volatility issues when disaggregated by both occupation and region.[2] Publicly available data on the ABS website only present regional data by the eight major occupation groups and is smoothed using an annual average for this reason.

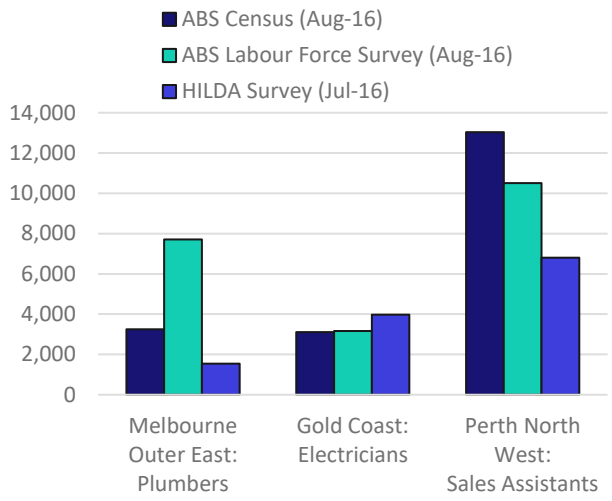
While all three potential sources mentioned above are imperfect due to their infrequency, sparsity or long time-lag between collection and release, the underlying estimates also differ. For example, the figure below shows some series record similar employment estimates across the three sources, and others record significant differences. This is to be expected given the different methodologies of each

[1] The sample size of the HILDA survey is approximately 17,000 which is too small to support the derivation of robust estimates at detailed levels of employment by region and occupation, numbering approximately 31,240 in total.

[2] The data are subject to significant volatility, large standard errors and a high number of missing values due to the relatively small sample size of the Labour Force Survey (50,000) relative to the number of estimates at detailed levels of region and occupation (31,240).

source. However, the differences highlight challenges in understanding labour market activity at granular levels and the need for more timely and frequent estimates.

Figure: Comparison of employment levels for selected occupations by region – ABS Census, ABS LFS and HILDA



3. The NSC's approach to developing NERO

The NSC's NERO is a monthly nowcast of employment in Australia by:

- Australian and New Zealand Standard Classification of Occupations (ANZSCO) (ABS, 2013) 4-digit occupation, and
- Statistical Area 4 (SA4) region (ABS, 2016).

The NERO estimates include approximately 355 occupations and 88 regions, equating to around 31,240 estimates in total.

The process undertaken to develop NERO including validation of the model's outputs is described below.

a) Sourcing relevant input data

Labour market data at detailed levels for Australia is hard to find, especially if it needs to be reasonably current. Many nowcasting attempts in the past have been made possible by a range of real-time data that is available at aggregate levels (such as financial market data, sentiment data, Google search trends data and the like). However, at more disaggregated levels – such as when examining regional and occupational components – the data are less readily available, particularly as the data need to be timely and frequent enough to support modelling.

The NSC assembled a range of data sources for testing in the model:

Source	Series
ABS Census of Population and Housing	Occupational employment by region (5 yearly)
ABS Labour Force Survey	Occupational employment nationally (quarterly)
	Occupational employment by region (quarterly)[3]
	Total employment by region (monthly)
NSC Internet Vacancy Index	Online job advertisements by region and occupation (monthly)
Burning Glass	Online job advertisements by region and occupation (daily)
DESE Jobactive program data	Jobactive job placements by occupation and region (fortnightly)
ABS Weekly Payroll Jobs	Weekly Payroll Jobs by Industry and region
Home Affairs	Visa holders by occupation and State/Territory (quarterly)
ABS National Accounts	Gross State Product (annual)

Further information on the data sources included in the model is provided in Appendix A.

b) Data cleaning and enhancement

Various treatments and transformations were made to the data so that it was fit for training and testing using machine learning techniques. These included:

- cross-checking release dates and reference periods to ensure that any data being used to predict a date in the past was only based on data released prior to that prediction date
- mapping all regional data to SA4 boundaries (using various geographical boundary concordances)
- aligning all industry-based data to an ANZSCO 4-digit occupational basis (using a concordance of industry to occupational employment from the 2016 Census)

[3] Custom, unpublished data for reasons stated in footnote 2.

- excluding series that were considered out of scope (such as Defence related occupations, not-further-defined occupations and other territories)[4]
- imputing missing values in the data where necessary using various imputation techniques based on the mean/median of the series and the surrounding data values
- smoothing the data using a Hodrick-Prescott filter,[5] (we also tested using a range of other smoothing techniques including Baxter King, Christiano Fitzgerald, Butterworth and several others).

For all input data sources, various features were created and tested as model inputs. This included using raw data, the change in raw data levels (weekly, fortnightly, monthly, quarterly, and annual) and lagged values.

c) Modelling process to generate initial estimates

Following the data cleaning and enhancement process, the NSC utilised a number of machine learning modelling techniques to develop predictions of employment by occupation and region.

Following standard machine learning practice for training, validation and testing, we split the data into two data sets:

1. **training and validation data set** (August 2015 to February 2020, excluding August 2016)
2. **testing data set** (August 2016, to enable an in-sample validating using the 2016 Census, and May 2020 to November 2020, to enable an out of sample validation)

The model is trained and tuned (developed) on the training and validation data set. Model performance testing is then completed by running the built models on the second testing data set to examine how the resultant predictions perform.

Three different but commonly utilised machine learning modelling approaches were deployed to generate predictions:

- *Random Forest* (Breiman 2001): This model uses a large number of "trees" which are developed independently of each other to allow for uncorrelated errors to improve performance. While some trees may be less accurate in some circumstances, many other trees will be more accurate – in effect, the trees protect each other from their individual errors. The final prediction is derived by taking the average prediction (or most common prediction) across all the "trees".
- *Gradient Boosting* (Friedman 2001): Similar to Random Forest, the Gradient Boosting model involves estimating "trees" which seek to explain the target variable. However, while under Random Forest each "tree" is built independently, Gradient Boosting builds one "tree" at a time, with each new "tree" seeking to improve on the shortcomings of the previous version of the model. This iterative tree building process continues until the learning algorithm is unable to develop new "trees" to explain the residuals.
- *Elastic net regression* (Zou and Hastie 2005): The elastic net regression is a common linear regression with an extra regularisation term. This penalises complex models and thus encourages smoother fitting.

These three main approaches form the main components of the machine learning approach. Notably each approach includes many sub-models. The three approaches can also be used to form a single combined estimate.

d) Combining multiple models into a single estimate

Once the Random Forest, Gradient Boosting and Elastic Net models were run, initial estimates were combined or stacked based on their hyper parameters, to produce a single, optimal set of nowcasts.

Building the stacked model involved taking the final output from each model as inputs to a linear regression (LR) model. The LR model was then trained to optimally combine the inputs, again utilising the training and validation and testing data sets. Once this LR model was optimised, we arrived at a single raw prediction of employment by region and occupation.

Generally, a stacked ensemble framework can obtain more accurate predictions, improved robustness and generalisability compared with the best individual model (Wolpert, 1992; Opitz and Machin, 1999). This approach was adopted in developing the NSC NERO estimates. It is considered this will provide a more stable and reliable model for the NSC in the medium to long term.

e) Adjusting outliers, smoothing and scaling to derive a final estimate

Once a single raw prediction was developed using the stacked ensemble model, an outlier adjustment, smoothing and scaling process was implemented to derive final estimates.

[4] Not further defined is a code used by the ABS to process incomplete, non-specific or imprecise responses. They were excluded from NERO as they were generally a low percentage of records. Defence-related occupations were also removed as it is a unique occupation series with poor data available.

[5] Smoothing using filters like the Hodrick-Prescott filter is commonly used in macroeconomics to extract a trend component from a time series.

Where the rate of change in the model's raw prediction deviates substantially from the national rate of change for Australia (as estimated using the ABS Labour Force Survey), an outlier adjustment process was completed. This ensures the nowcasting estimates of smaller series, in particular, which are often volatile, remain broadly consistent with the national trends. A minimum of 10 people employed was also applied to all series to help ensure confidentiality.

Once the outlier adjustment process was completed, the estimates were smoothed using the Hodrick-Prescott Filter to provide trend estimates. The estimates were then scaled to ensure broad consistency with the known national trends identified through the ABS Labour Force Survey. This involved broadly scaling to ABS Labour Force Survey estimates for total employment in each region, as well as the estimates for employment by occupation for Australia. This ensures that the NSC nowcasting predictions are broadly consistent with existing estimates of total employment for each region and occupation.

Once this process was completed, a final smoothed nowcasting estimate was derived.

4. Validation of the NSC's NERO model

As mentioned earlier in this paper, existing data on regional employment by occupation in Australia is limited, making validation of the NERO model challenging – particularly for smaller level series.

To address the challenges, the NSC developed a number of potential validation measures including:

- for larger series, examining the accuracy of the NSC nowcast estimates against a smoothed version of the ABS Labour Force Survey custom data on occupational employment by region (quarterly), and
- for smaller series, examining the accuracy of the NSC nowcast estimates against the outcomes from the 2016 ABS Census of Population and Housing.

Together, these two sources – although imperfect – provide an appropriate source of data to validate the NSC NERO model against.

Performance of the model was evaluated using three measures – the MAPE, WAPE and RMSE. For all three metrics, the smaller the value, the better the model. These metrics can be derived as:

- **Mean Absolute Percentage Error** (MAPE): a measure of how much each prediction has 'missed' by in percentage terms:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i - \hat{x}_i}{x_i}\right|$$

- **Weighted Absolute Percentage Error** (WAPE): a weighted measure that penalises errors for larger series:

$$WAPE = \frac{\sum_{i=1}^{N}|x_i - \hat{x}_i|}{\sum_{i=1}^{N}|x_i|}$$

- **Root Mean Square Error** (RMSE): a measure which penalises larger errors:

$$RMSE = \sqrt{\frac{\sum_{i=i}^{N}(x_i - \hat{x}_i)^2}{N}}$$

where x_i is the actual value, $\hat{x}_i$ is the predicted value and N is the number of data points.

While these measures and the above stated data sources provide a method to measure model performance, it still does not shed light on what should be considered adequate or sufficient performance. Since the NSC NERO is one of the first attempts to create new data at disaggregated levels using big data and machine learning techniques, there is limited precedence in the literature that can be used to understand the performance. With this in mind, the NSC constructed a simple model for the benchmarking purpose. This simple 'benchmark' model uses a smoothed version of the ABS Labour Force Survey data to predict the next value in the time series.

The table below provides an overview of performance metrics for the NSC NERO model, including for each of the contributing models (Random Forest, Gradient Boosting and Elastic Net) as well as the benchmark model. The performance is measured on the testing dataset.

Model	MAPE	WAPE	RMSE
Benchmark	22%	18%	305
Random Forest	16%	13%	231
Gradient Boosting	19%	13%	237
Elastic Net	20%	13%	219
NERO (stacked)	**20%**	**14%**	**246**

The table above shows that the models outperformed the benchmark. Three individual model types and the stacked model perform similarly, with similar metrics recorded across all three models and performance measures (MAPE, WAPE and RMSE). While the stacked final NERO model performs slightly worse than some of the individual models, this model is still considered preferable at this stage given

existing evidence suggests this model is likely to be more reliable and stable than a given single model in the medium to long term.

This level of performance was considered appropriate and usable, particularly given two of the three periods the model is measured against include the impacts of COVID-19, which was extremely difficult to predict. The table below shows how the performance of the model expectedly declines slightly during the height of COVID-19 in Australia in 2020.

Period	Description	WAPE (stacked model)
Aug-16	Prior to COVID-19	11%
May-20	COVID-19 downturn	13%
Aug-20	COVID-19 recovery	17%
Overall		**14%**

It is envisaged that with additional data incorporated into the model (such as bank data) and a more stable prediction period, the errors will decline. Additional breakdowns of model performance are also available in Appendix A.

5. Overview of model outputs

The NSC NERO data provides new estimates of regional employment by occupation on a monthly basis for 31,240 series. The figures below provide example outputs to help visualise the data.

Figure: NSC NERO Employment estimate for Education Aides in Melbourne – North East

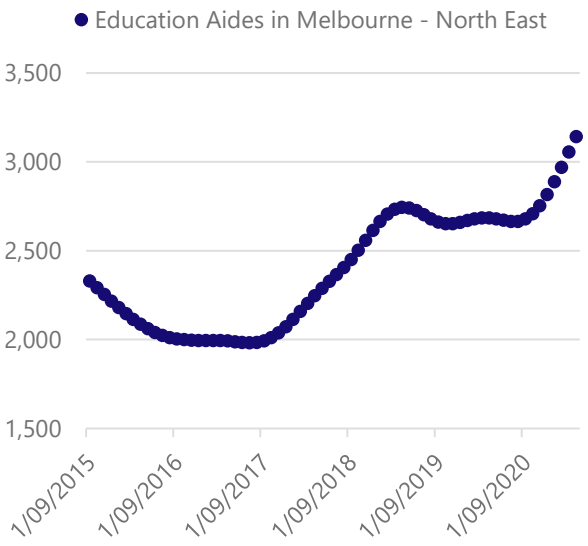


Figure: NSC NERO Employment estimate for Software and Applications Programmers in Brisbane – South

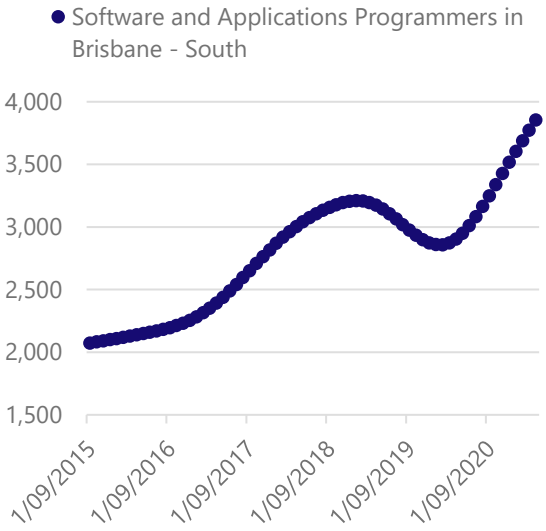
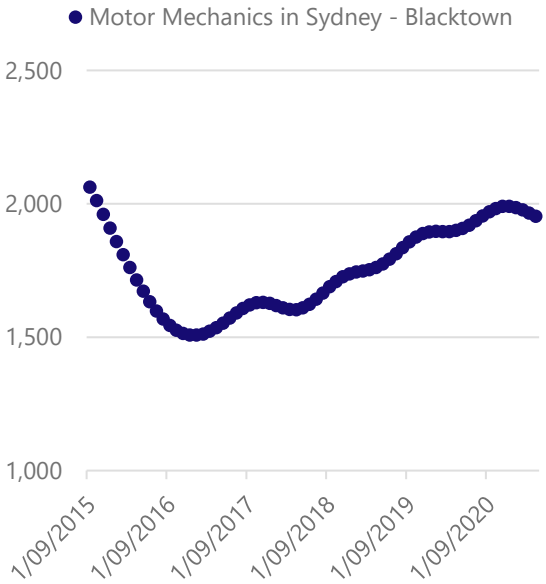


Figure: NSC NERO Employment estimate for Motor Mechanics in Sydney – Blacktown



The outputs enable the identification of the largest employing regions for each occupation. For example, the following table shows the 10 largest employing regions for Aged and Disabled Carers.

Employment of Aged and Disabled Carers by Region (SA4)	Employment (NSC NERO) – April 2021	5-yr change
Gold Coast	9,043	63%
Melbourne - West	8,373	63%
Melbourne - South East	6,725	4%
Perth - North West	6,711	41%
Adelaide - North	6,596	45%
Adelaide - South	6,578	39%
Perth - South East	5,840	83%
Melbourne - Outer East	5,701	22%
Wide Bay QLD	5,208	80%
Perth - South West	5,082	70%
Sunshine Coast	4,868	73%
Melbourne - North East	4,654	84%
Capital Region NSW	4,027	41%

From a regional perspective, the estimates enable the identification of the top occupations in each region, as well as the fastest growing occupations in the region as shown, for example, in the following table.

Employment in Illawarra (NSW) by occupation	Employment (NSC NERO) – April 2021	5-yr change
Sales Assistants (General)	6,623	2%
General Clerks	6,122	55%
Registered Nurses	4,762	32%
Aged and Disabled Carers	3,325	25%
Electricians	2,834	50%
Primary School Teachers	2,758	67%
Metal Fitters and Machinists	2,613	18%
Carpenters and Joiners	2,592	13%
Office Managers	2,537	40%
Retail Managers	2,336	7%

Further outputs can be examined via the data dashboard on the NSC's website where the data can also be downloaded.

6. Potential use cases for the new NERO outputs

The NSC NERO outputs will provide a range of new ways to examine the labour market. This includes enabling the timely analysis of occupational employment at a regional level for the first time (including the top employing occupations and the fastest growing occupations), while also supporting the identification of turning points in the labour market at regional levels.

The successful production of reliable nowcasting estimates of employment by detailed occupation and region may have a range of potential analytical and policy applications. Examples relevant to the DESE portfolio include:

- supporting employment service providers and training providers to better target their service offerings to the jobs in demand in their region,
- more effectively targeting policy responses to local conditions, including policy responses that seek to address structural adjustment issues within industries and regions, and
- using the estimates to more effectively account for regional differences when evaluating labour market programs and setting performance benchmarks for service providers.

The NSC NERO estimates for the 31,240 series of occupational employment by region will be published monthly on the NSC website going forward, including a data dashboard and data download functionality to support users to conduct their own analysis of the data. There are however, two important caveats to the use of the data:

- The estimates are currently experimental in nature, and may be revised by the NSC in late 2021 or early 2022 in response to feedback from stakeholders and the inclusion of additional data sources into the model
- The primary purpose of the NERO data is to complement existing sources of data and information on employment by occupation and region. The NERO data should therefore be used in conjunction with sources from the ABS and others, rather than as a standalone source.

Conclusion

This paper outlines one of the first attempts in the world to nowcast the labour market at a detailed disaggregated level. The analysis extends upon existing literature on nowcasting, which primarily focusses on nowcasting economy-wide aggregates such as national GDP or unemployment rates.

This paper shows that combining a range of traditional and real-time data sources can produce a useful and timely indicator of employment by occupation and region in Australia.

Future performance of the NSC NERO model could be improved through the incorporation of more sources of timely and disaggregated data (such as bank transaction or accounting data) and through further model training and validation using data from future censuses (such as the forthcoming 2021 Australian Census of Population and Housing).

One of the primary purposes of publishing NERO is to identify new use cases that may not be evident to the NSC. The sharing of intelligence aims to enrich Australia's capacity to better understand and adapt to the changing labour market.

The NSC welcomes all feedback regarding potential use cases and model improvements via:

nowcasting@skillscommission.gov.au

References

ABS (2013) *Australian and New Zealand Standard Classification of Occupations (Cat No 1220.0)* Version 1.3 https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1220.0Chapter22013,%20Version%201.3

ABS (2016) *Australian Statistical Geography Standard* (ASGS), Volume 1 - Main Structure and Greater Capital City Statistical Areas (Cat No 1270.0.55.001), July, https://www.abs.gov.au/ausstats/abs@.nsf/mf/1270.0.55.001

Bok B, D Caratelli, D Giannone, A Sbordone, and A Tambalotti (2017) *Macroeconomic Nowcasting and Forecasting with Big Data*, Federal Reserve Bank of New York Staff Reports, no. 830, November, https://www.newyorkfed.org/research/staff_reports/sr830

Breiman L (2001) *Random Forests*, Machine Learning, 45, October, https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf.

Dawson N, M Rizoiu, B Johnston and M Williams (2020), *Predicting labour shortages from labour demand and labour supply data: A machine-learning approach*, manuscript submitted for review to the 21st ACM Conference on Economics and Computation, 6 April, https://www.oitcinterfor.org/sites/default/files/file_publicacion/Predicting_Labor-Shortages-MachineLearningApproach.pdf.

Friedman, J (2001), *Greedy Function Approximation: A Gradient Boosting Machine*, IMS 1999 Reitz Lecture, https://www.jstor.org/stable/2699986#metadata_info_tab_contents

Higgins, P (2014) *GDPNow: A model for GDP 'Nowcasting'*, Federal Reserve Bank of Atlanta Working Paper 2014-7, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2580350.

Kindberg-Hanlon G, A Sokol (2018) *Gauging the globe: the Bank's approach to nowcasting world GDP*, Bank of England Quarterly Bulletin, Q3, https://www.bankofengland.co.uk/quarterly-bulletin/2018/2018-q3/gauging-the-globe-the-banks-approach-to-nowcasting-world-gdp

Moriwaki, D (2020) *Nowcasting Unemployment Rates with Smartphone GPS Data*, Lecture Notes in Computer Science Book Series, LNCS Volume 11889, January, https://link.springer.com/chapter/10.1007/978-3-030-38081-6_3.

Nguyen K and G La Caga (2020) *Start Spreading the News: News Sentiment and Economic Activity in Australia*, Reserve Bank of Australia Research Discussion Paper RDP 2020-08, https://www.rba.gov.au/publications/rdp/2020/2020-08.html.

OECD (2017), *Nowcasting Trade in Value Added*, http://www.oecd.org/std/its/tiva-nowcast-methodology.pdf.

Opitz, D and R Maclin (1999), *Popular ensemble methods: An empirical study*, Journal of Artificial Intelligence Research, Volume 11, August, https://jair.org/index.php/jair/article/view/10239/24370.

Varian, H (2014) *Big Data: New Tricks for Econometrics*, Journal of Economic Perspectives, Volume 28, Number 2, Spring, https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3.

Wolpert, D (1992), *Stacked generalization Neural Networks*, Neural Networks, December, https://www.researchgate.net/publication/222467943_Stacked_Generalization.

Zou, H and T Hastie (2005) *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society, 67, part 2, pp. 301-320, https://rss.onlinelibrary.wiley.com/journal/1467985X

Appendix A: Additional information

Detailed list of data sources tested in the development of the NSC NERO model

Source	Series	Level	Regional level	Start date[6]	Frequency	Access	Comment
ABS – Census	Occupational employment by region	4-digit ANZSCO	SA4 region	TBC	Every five years	Via subscription	The most reliable existing estimate of employment for smaller series which are not typically captured by the ABS LFS or HILDA Survey.[7]
ABS Labour Force Survey	Occupational employment nationally	4-digit ANZSCO	State & Territory	Aug 1986	Quarterly	Publicly available	
	Occupational employment by region	4-digit ANZSCO	SA4 region	Feb 2001	Quarterly	Via subscription (custom data request)	Subject to significantly volatility, large standard errors and a high number of missing values.
	Total employment by region	Total employment	SA4 region	Oct 1998	Monthly	Publicly available	
NSC – Internet Vacancy Index	Online job advertisements by region and occupation	4-digit ANZSCO	IVI regions	Mar 2010	Monthly	Publicly available	As the data is based on IVI regions, trends at the SA4 level will need to be inferred through a concordance process.
Burning Glass	Online job advertisements by region and occupation	4-digit ANZSCO	SA4 region	Jan 2013	Daily	Via subscription	Does not have the same breadth of coverage as the NSC IVI, although it is more timely/frequent.
DESE – Jobactive program data	Jobactive job placements by occupation and region	4-digit ANZSCO	SA4 region	Jul 2015	Fortnightly	Government program data	Remote areas are not captured in this data as the jobactive program does not operate in remote areas.[8]
ABS – Weekly Payroll Jobs	Weekly Payroll Jobs by Industry and region	Total employment 1-digit and 2-digit ANZSIC	SA4 region State & Territory National	Jan 2020	Weekly	Publicly available	As a relatively new series, caution must be exercised in utilising this data. A separate model that utilises this data may be required.
Home Affairs	Visa holders by occupation and State/Territory	4-digit ANZSCO	State & Territory	Sep 2010	Quarterly	Publicly available	
ABS – National Accounts	Gross State Product (GSP)	1-digit ANZSIC	State & Territory	Jun 1990	Annual	Publicly available	The occupational impacts of economic activity by industry will need to be inferred through a concordance process.

[6] Start date indicates availability on a consistent time series basis

[7] The ABS Jobs in Australia series may potentially provide a detailed occupation by region picture using tax data on an annual basis.

[8] Placements are recorded by jobactive providers in the Employment Services System for job seekers on their caseload. Not all occupations where a job-seeker starts a new job are necessarily recorded as a placement, and placements are not recorded for participants in digital services. In response to the large increase to the jobactive caseload during the COVID-19 pandemic, Online Engagement Services was expanded leading to a significant change in the percentage of the jobactive participants in digital services. This means there is a break in series from April 2020 onwards.

Additional information on the modelling methodologies

The following section provides further information regarding the machine learning and modelling approaches utilised to produce the NSC NERO estimates, consistent with the existing literature available on these approaches.

Gradient boosting

Gradient boosting was introduced by Friedman (2001). It is an ensemble method that can combine several weak learners into a strong learner as:

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i)$$

where $f_k(.)$ is a weak learner and K is the number of weak learners, combined to be become a strong learner. Given a training dataset $D = \{(y_i, x_i)\}(|D| = n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$, one would like to find a stronger learner, which finds the optimal parameters by minimizing the loss function $\emptyset$, as shown in:

$$\mathcal{L}(\emptyset) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Here l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i. The second term Ω penalizes the complexity of the model. The additional regularization term helps to smooth the final learnt weights to avoid over-fitting. The tree ensemble model used in XGBoost is trained in an additive manner until stopping criteria (e.g., the number of boosting iterations, early stopping rounds and so on) are satisfied.

The basic procedure of boosting is described in pseudocode below:

Set uniform sample weights.

for each base learner (weak learner) do

Train base learner with weighted samples.

Test base learner on all samples.

Set learner weight proportional to weighted error.

Set sample weights based on ensemble predictions.

Weighted average all base learners as the final model.

Random Forest

The random forest (RF) algorithm, proposed by L. Breiman in 2001, has been extremely successful as a general-purpose classification and regression method. The approach, which combines several randomised decision trees and aggregates their predictions by averaging, has shown excellent performance in many applications. The idea in random forest algorithm is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much.

The basic procedure of Random Forest is described in pseudocode below:

for i = 1 to N do

Randomly select k features from total features.

Randomly select d samples from total learning samples.

Build a tree with selected k features and selected d samples.

Average all N trees as the final model.

Elastic Net

The elastic net method is a recent development in regression. It can be understood as a conventional regression with a penalty term. Given a training dataset with n observations and p predictors. Let $y = (y_1, \dots, y_n)^T$ be the response and $X = [x_1 | \dots | x_p]$ be the model matrix, where $x_j = (x_{1j}, \dots, x_{nj})^T, j = 1, \dots, p$ are the predictors. Elastic Net regression attempts to minimise the residual sum of squares plus some penalty term.

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda[(1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1]$$

Here, $\|\beta\|_1$ is the Lasso penalty called the l_1 norm:

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

Similarly, $\|\beta\|_2$ is the Ridge penalty called the l_2 or Euclidean norm:

$$\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$$

Additional model performance metrics and breakdowns

Model performance of various breakdowns of the series were investigated to find potential weak spots in the model that could be improved in the next iteration.

Series that were stable or only exhibited small increases or decreases tended to have better performance. The series with the largest declines exhibited the largest errors.

Table: Summary of model performance by trend

Series	Annual change of smoothed LFS	WAPE (stacked model)
Large increase	Greater than 15%	13%
Small increase	Between 2.5% and 15%	10%
Stable	Between -2.5% and 2.5%	8%
Small decrease	Between -10% and -2.5%	11%
Large Decrease	Less than -10%	18%
Overall		**14%**

The model performance is consistent across states and territories.

Table: Summary of model performance by State

Series	WAPE (stacked model)
Australian Capital Territory	13%
New South Wales	14%
Northern Territory	15%
Queensland	14%
South Australia	13%
Tasmania	13%
Victoria	14%
Western Australia	14%
Overall	14%

A breakdown by the ANZSCO 1-digit occupation also revealed consistent model performance.

Table: Model performance by ANZSCO 1-digit

Series	WAPE (stacked model)
Managers	14%
Professionals	13%
Technicians & Trades Workers	14%
Community & Personal Service Workers	15%
South Clerical & Administrative Workers	14%
Sales Workers	13%
Machinery Operators & Drivers	14%
Labourers	14%
Overall	14%

The models performed best for the largest series. High errors are present in the smallest group of between zero and 100 employed due to a minimum value of 10 being applied to all predictions.

Table: Summary of model performance by number of people employed

Series	WAPE (stacked model)
Between 0 and 100	88%
Between 101 and 500	14%
Between 501 and 1000	14%
Between 1001 and 5000	13%
Between 5001 and more	11%
Overall	**14%**

For more information
on this report, contact:

Nowcasting and Economic Modelling Section
nowcasting@skillscommission.gov.au

nationalskillscommission.gov.au

Australian Government

National Skills Commission

ESE20–0203