Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Evaluation of the best M4 competition methods for small area population forecasting

Tom Wilson [*], Irina Grossman, Jeromey Temple

The University of Melbourne, Melbourne School of Population and Global Health, The University of Melbourne, 207 Bouverie St, Melbourne, Vic 3010, Australia

ARTICLE INFO

Keywords:
Population forecast
Forecast error
Small area
M4 competition
Australia
New Zealand

ABSTRACT

The ‘M4’ forecasting competition results were featured recently in a special issue of the *International Journal of Forecasting* and included projections for demographic time series. We sought to investigate whether the best M4 methods could improve the accuracy of small area population forecasts, which generally suffer from much higher forecast errors than regions with larger populations. The aim of this study was to apply the top ten M4 forecasting methods to produce 5- and 10-year forecasts of small area total populations using historical datasets from Australia and New Zealand. Forecasts were compared against the actual population numbers and forecasts from two simple benchmark models. The M4 methods were found to perform relatively well compared to our benchmarks. In the light of these findings, we discuss possible future directions for small area population forecasting research.

1. Introduction

Small area population forecasts are used by government and business to plan for future childcare demand, housing, new schools, aged care services, health facilities, emergency service provision, power and water demand, transport use, and electoral redistricting, amongst other uses. They inform decisions that can involve millions of dollars in investments in a local area (or not) and, therefore, have important consequences for local service provision. Unfortunately, small area population forecasts do not enjoy a glowing record of accuracy. They are far more error-prone than national or state population forecasts, with errors tending to increase as the population size of a small area decreases, especially once the population falls below about 10,000 (Wilson, Brokensha, Rowe, & Simpson, 2018). Partly this is due to data limitations, including short time series at the small area scale, poor quality data, and random noise in small populations making underlying demographic patterns difficult to discern. But it is also probably due in part to limited research. The amount of research effort in trying to improve the accuracy and utility of small area forecasts is dwarfed by research on national-scale population modelling - yet the need for research to improve small area forecasting is arguably greater.

Over the last decade, relatively little research on small area population forecasting methods has been published. A key finding (though not new to the forecasting literature generally) is that combining or averaging forecasts can reduce errors. This includes averaging several forecasts of the total population as well as constraining cohort-component forecasts of the population by age groups and sex to independent total populations forecasts (e.g., Reinhold & Thomsen, 2015; Wilson, 2014, 2015, 2016). Incorporating spatial relationships between small areas has also been found to improve accuracy (e.g., Baker, Alcántara, Ruan, Watkins, & Vasan, 2014), though models incorporating a range of socio-economic variables have been less successful than expected (e.g., Chi & Voss,

[*] Corresponding author.
E-mail addresses: wilson.t1@unimelb.edu.au (T. Wilson), irina.grossman@unimelb.edu.au (I. Grossman), jeromey.temple@unimelb.edu.au (J. Temple).

https://doi.org/10.1016/j.ijforecast.2021.09.005

2011). Researchers preparing population forecasts to assess future climate change effects often 'downscale' existing national or regional forecasts to grid cells using approaches ranging from simple extrapolation to machine learning forecasts of land use development (e.g., Chen, Li, Huang, Luo, & Gao, 2020; Hachadoorian, Gaffin, & Engelman, 2011). However, the accuracy of these downscaled forecasts remains largely unknown.

Although newer approaches involving machine learning methods, including neural networks, are just beginning to be used for subnational population forecasting (e.g., Riiman, Wilson, Milewicz, & Pirkelbauer, 2019; Weber, 2020), there is potential to engage more with new methods and technologies, and forecasting developments outside demography. One avenue for pursuing this engagement is via the M Competitions. The M Competitions were created to support the advancement of the research and practice of forecasting by challenging researchers to develop methods which can accurately forecast time series from multiple application domains with different data frequencies (Makridakis, Spiliotis, & Assimakopoulos, 2020a). Papers describing these competitions have been amongst the most impactful research articles in forecasting and have served to '…bridge the gap between theory and practice' (Petropoulos & Makridakis, 2020, p. 3). Perhaps nowhere is this gap greater than between the fast-changing world of forecasting research and the practical world of applied demography, particularly at the small area scale. The recent announcement of the M4 forecasting competition outcomes (Makridakis et al., 2020a) offers an opportunity to apply the best methods from that competition to small area population datasets. Given that demography was one of the application domains of the M4 competition (Makridakis et al., 2020a), we were very interested to see whether the M4 competition methods applied to small area populations would produce more accurate forecasts than current methods used by demographers.

This paper reports on an evaluation of the top ten best-performing methods in the M4 competition applied to forecast small area population totals (i.e., without any breakdown by age group and sex). Using small area population estimates for Australia and New Zealand, we prepared 'forecasts' for historical periods for 5 and 10 years ahead, which are typical horizons for many local area planning purposes. We evaluated forecast performance by comparing them with published population estimates and the forecasts from two simple benchmark models. The evaluation included a breakdown by population size category, growth rate, and position in the settlement hierarchy to determine if some types of methods worked best in particular types of areas. If this turns out to be the case, then forecasts could be produced using a composite approach in which different methods are applied to different area types.

In Section 2, we describe the population data, fitting periods, forecast horizons, the ten M4 competition methods, two benchmark models, and error measures. The results of the evaluation are presented in the next section, whilst Section 4 consists of a discussion which includes a summary of key findings, a description of the complex pattern of results, and avenues for further research.

2. Material and methods

2.1. Population data

We obtained mid-year estimated resident population (ERP) totals for SA2 areas of Australia for the period 1991 to 2016 from the Australian Bureau of Statistics (ABS, 2017), and similar ERP totals for SA2 areas of New Zealand for the period 1996 to 2020 from Statistics New Zealand (2020a). SA2 areas comprise a key small spatial unit in the official statistical geographies of both countries, and population data and forecasts are regularly prepared for these areas. SA2 areas in Australia in 2016 had a median population of 9,681 with 95% lying within the range 2559 to 29,279; in New Zealand, SA2 areas are smaller, with the median population in 2020 being 2340, and the 95% range 400 to 4,800.

The selected periods of SA2 area populations were the longest time series available on a consistent set of geographical boundaries. The Australian ERPs are all based on SA2 geographical boundaries existing in 2011 whilst those for New Zealand are based on boundaries defined in 2020. A few SA2 areas in each country were omitted from the analysis because they had populations under 100 in one or more of the fitting periods of the models tested in this study. These SA2s were all combined into a single 'remainder' area, giving a total of 2067 SA2 areas for Australia and 2054 for New Zealand. The ERP datasets used in this study are available in the supplementary material.

National population forecasts were required as data inputs or constraints to the small area forecasts. We used the main series forecasts produced by the ABS and Statistics New Zealand closest in date to the jump-off years of the forecasts (ABS, 2008, 2013a; Statistics New Zealand, 2009, 2014).

2.2. Forecasting methods

The top ten forecasting methods of the M4 competition are summarized in Table 1. The code to run each of the methods is available from the M4 competition GitHub repository (https://github.com/Mcompetitions/M4-methods). Most of the code is in R, except for the Smyl (2020) method which is in C++, and the Doornik, Castle, and Hendry (2020) method which was created in Ox version 7.20 language (Doornik, 2013; Doornik & Ooms, 2006). We used Visual Studio 2019, OxEdit 7 (Doornik, 2013; Doornik & Ooms, 2006), R version 4.0.2 and RStudio Version 1.3.1093, and the latest associated R toolboxes available. Data up to the jump-off year were used as input to the models. We endeavoured to apply the M4 methods with minimal modifications to their competition submissions, such that where possible we only modified the parts of the code related to the data input, the use of only yearly data (as opposed to the six frequencies in the original M4 dataset), and the forecast horizons. The exception is the Smyl method, which required some adaptation, as described below.

The time series for the M4 competition are significantly longer than those available to us for our small

Table 1
Summary of the top ten M4 competition forecasting methods.
Source: Makridakis et al. (2020a) and references cited in the table.

Reference	M4 rank	Method summary
Smyl (2020)	1	A hybrid model incorporating both a recurrent neural network and exponential smoothing formulae. The gradient descent method is used to fit the parameters of both the statistical and neural network aspects of the method.
Montero-Manso, Athanasopoulos, Hyndman, and Talagala (2020)	2	A combination of 9 forecasting methods; weights are calculated using a gradient tree boosting-based learning model. The models include both statistical methods and a neural network: ARIMA, ETS, TBATS, STLM-AR, RW-DRIFT, THETAF, Naive, Seasonal Naïve, and NNETAR.
Pawlikowski and Chorowska (2020)	3	The method for the forecasting of yearly data involves clustering the time series in the M4 dataset based on similar trends and then selecting a set of methods for each cluster. This set is chosen from Naive, exponential smoothing, theta, ARIMA, and linear regression models. A rolling origin evaluation is then used to determine the performance of these models and assign weights to each of them.
Jaganathan and Prakash (2020)	4	An ensemble of methods including ETS, ARIMA, damped ETS, Naïve/SNaive, MAPA, theta, and hybrid theta. A simple median operator is used to combine the methods. The authors' submission to the M4 competition also included ForecastPro (Business Forecast Systems, Inc., 2018), commercial software that was not available to us and was excluded from our implementation of this method.
Fiorucci and Louzada (2020)	5	The GROEC method is a combination of the DOTM, OTM, ETS, and ARIMA statistical models where weights are assigned using a cross-validation scheme.
Petropoulos and Svetunkov (2020)	6	The medians of the following univariate models are used to create their forecasts: automatic ARIMA, complex exponential smoothing, exponential smoothing, and dynamic optimized theta method.
Shaub (2020)	7	The authors used the forecastHybrid package (Shaub & Ellis, 2018) to create a forecast using an ensemble of methods which included: ARIMA, THETAF, and TBATS. A simple average of these three methods was used.
Legaki and Koutsouri (2018)	8	This method uses a Theta-Box–Cox method. This statistical method involves several steps, beginning with a seasonality test and deseasonalization where required. The Box–Cox transformation is then applied, followed by forecasts using the theta method. A reverse Box–Cox transformation is then applied, and the data re-seasonalized, if required.
Doornik et al. (2020)	9	This forecasting method first creates two separate forecasts using the rho (adaptive autoregressive model) and delta (dampened trend from growth rates) methods. These two forecasts are then averaged. This average is then calibrated; it is treated like observed data and fitted with an autoregressive model which is then used to produce the final output.
Pedregal, Trapero, Villegas, and Madrigal (2018)	10	For yearly data, this submission involved the application of the Theta-4-ARMA method. Forecasts are first created using the Theta-4 method. However, the Theta-4 method often produces auto-correlated residuals. An ARMA model is then used for these residuals.

Notes: Rank is based on the performance of the point forecasts for the M4 competition. It was determined using the overall weighted average of two measures of accuracy, the scaled mean absolute percentage error (SMAPE) and the mean absolute scaled error (MASE) (Makridakis et al., 2020a). Many of the methods relied on the Forecast R package in R (Hyndman et al., 2018), including those using the ARIMA, automated exponential smoothing algorithm (ETS), neural network time series forecasts (NNETAR), exponential smoothing state-space model with Box-Cox transformation (TBATS), time series decomposition using the STL method with an autoregressive model (STLM-AR), random walk with drift (RW-DRIFT), theta method (THETAF), Naive, and Seasonal Naïve. The theta model (DOTM) and optimized theta model (OTM) models were implemented with the forecTheta package (Fiorucci, Louzada, & Yiqi, 2016). Please refer to the documentation associated with these R packages for information on the individual models.

area population forecasts. The methods developed for the competition, particularly those using machine learning, are generally better suited for longer time series. For this reason, we used the full-time series available up to the jump-off year for our forecasts (data from 1991 for Australia, and from 1996 for New Zealand). This contrasts with the benchmark methods which only require 10 years of base period data and use only the base and launch years to fit their models; this allows them to be used for the common case where small areas only have decennial census counts available. This base period length has been evaluated and found to be effective previously (Rayer & Smith, 2010). We evaluated the impact of extending the base periods to the maximum available time series for the forecasts reported here and found no benefit.

The winning method of the M4 competition was created by Slawek Smyl (2020) using C++ with the DyNet library (Neubig, Dyer, Goldberg, Matthews, Ammar, Anastasopoulos, Ballesteros, Chiang, Clothiaux, & Cohn, 2017). Smyl's method uses a hierarchical approach which incorporates both the long short-term memory (LSTM) neural network, which supports learning across multiple time series, and the exponential smoothing model (ES), which allows local features at the level of individual areas to be taken into consideration. The method is a hybrid of machine learning and statistical models - rather than a simple combination - as it fits the parameters of both simultaneously using the same gradient descent method. Rolling input and output windows are used during preprocessing, whereby subsamples of the training data are used to train the models. The output window used is the same size as the forecasting horizon; for example, if the output window is set to 5 years during training, the model will produce a forecast with a 5-year horizon. The input window should be similar to the forecasting horizon for optimal forecasts. We kept the input size to the default of 4 years for the 5-year horizon forecasts; this was the input window size selected for yearly data in the M4 competition, where the forecast horizon was 6 years. Ideally, we could have increased the window size for the 10-year horizon forecasts, but our time series are too short. This is standard in small area demography. Initially, we produced 10-year forecasts using an input window width of 4 for the 10-year 2006-based Australia forecasts, and 3 for the 2010-based New Zealand forecasts. This produced poor results, as expected. Rather than leave the method out of the 10-year evaluations, we employed an iterative approach where we used a 4-year input window to forecast 5 years ahead and then appended this forecast to the training data and created another 5-year forecast using the appended data. This produced significantly better results than using reduced window widths. The original Smyl (2017) method utilizes a time series application domain as an external variable. We evaluated using an area's remoteness as an external variable. The modified method incorporating a remoteness category is labelled 'Smyl (r-c)'. Smyl (2017) suggested that 6–8 independent runs of the model would be suitable for an ensemble, so we ran 8 independent runs and then used the provided R script to average them into the final forecast. Separate models were trained for the Australian and New Zealand forecasts for the Smyl (2020) and Montero-Manso et al. (2020) methods.

The method in second place overall in the M4 competition was devised by Montero-Manso et al. (2020). This uses a combination of 9 forecasting methods, including both statistical methods and a simple neural net, with a gradient tree boosting-based learning model to assign weights to each of the methods. A temporal holdout approach is used for training the learning model to minimize forecast error. This involves withholding the last h values from each time series, where h should be the length of the required forecast horizon, and then applying each of the forecasting methods to the remaining section of the time series. The length of the remainder must be >7, otherwise, a smaller h must be used. We used $h = 5$ for the 2015-based (New Zealand) and 2011-based (Australia) 5-year forecasts, but for the 10-year forecasts, h needed to be reduced to 8 for Australia and 7 for New Zealand as the time series are too short to equal the full forecast horizon.

The remaining M4 methods are either statistical methods or combinations and can generally be run without needing any hyperparameter tuning. These methods are briefly described in Table 1. However, to support the replication of our results, we mention here some code-related changes that were required beyond those relating to the input datasets, changes to frequency (some methods were created with the expectation of multiple data frequencies), and the length of the forecasting horizon. Jaganathan and Prakash's (2020) method for yearly forecasts is based on an ensemble of methods. For the M4 competition, this ensemble included a forecast created using ForecastPro (Business Forecast Systems, Inc., 2018), which is commercial software unavailable to us. We therefore excluded it from our implementation of this method. In addition, their method makes use of the pbmcapply package in R (Kuang, Kong, & Napolitano, 2019) which supports parallelization of code, but not on Windows machines, which were used to create the forecasts here. The code was therefore changed for single-core processing. Fiorucci and Louzada's (2020) method required a minor modification which changed their p parameter from 6 years to 5 years in their *runModels* function, otherwise, the same value was produced for each of the years in the forecast period. Modifications were also made to Shaub's method such that it only required data from one data frequency (yearly), and this yearly data was fed into the forecasting method one time series at a time. The Doornik et al. (2020) method required installation of Ox version 7.20, and we used the OxEdit 7 text editor to run the code (Doornik, 2013; Doornik & Ooms, 2006).

The forecasts from the ten M4 methods are compared to those produced by two benchmark models. The first is a linear/exponential model (LIN/EXP) in which a small area's population is forecast by linear extrapolation if growth over the past decade has been positive, and by exponential extrapolation if it has been negative (Wilson, 2015). The exponential extrapolation ensures that populations experiencing a sharp decline over the base period are not forecast to become negative in the long run. The data requirements consist simply of small area population totals for the jump-off year and 10 years earlier. They are used to calculate annual average numerical growth (linear) or annual average growth rates (exponential) for the 10-year base period. The jump-off year is a term commonly used in demography for the last year in the base period (or training dataset).

The second benchmark model is a constant share of population and a variable share of growth (CSP-VSG) averaged model, where the shares of population and population growth are of the national population (Wilson, 2015). The constant share of population forecast is calculated by multiplying each small area's share of the national jump-off year population by a separate forecast of the national population. This implies that each small area is forecast to grow at the same rate as the national population. The variable share of growth forecast is initially calculated by applying a LIN/EXP model to each small area, and then adjusting growth using the plus-minus method (Smith, Tayman, & Swanson, 2013) to ensure that forecast population growth across all small areas matches national forecast growth. The final forecast is the mean of the two-component models' forecasts. The data requirements are exactly the same as for the LIN/EXP model.

2.3. Retrospective forecasts

For the top ten M4 methods and the two benchmark models, we produced forecasts for two jump-off years. For Australia, we created 2006-based and 2011-based forecasts out to 2016. For New Zealand, we created 2010-based and 2015-based forecasts out to 2020. The base periods for the benchmark models were the decade prior to each jump-off year, whilst for the M4 methods, the base periods extended from the start of each small area population time series to the jump-off year (for Australia, 1991–2006 and 1991–2011; for New Zealand, 1996–2010 and 1996–2015).

For all of the above forecasts, two versions were created:

(1) Unconstrained - forecasts as directly output by each of the models.
(2) Constrained - forecasts from each model constrained to national population forecasts.

Small area population forecasts produced by practising demographers often have to be consistent with forecasts produced at higher geographical scales. We applied a simple scaling factor to all small areas which is the national population forecast divided by the total unconstrained forecast for all small areas.

The Prestons area in New Zealand suffered runaway growth in the 2015-based forecasts from the Montero-Manso et al. and Shaub methods, with forecasted populations exceeding 1 million and 43 million, respectively. These errors would have been identified by a practitioner during the forecast preparation phase or during the review of the draft forecasts. Because the errors were readily identifiable and they significantly affected the forecast results, they were replaced by those of the LIN/EXP model. The LIN/EXP model was chosen because it is easier to calculate than the CSP-VSG model.

2.4. Evaluation measures

The main error measure used to evaluate the forecasts is median absolute percentage error (MedAPE). Percentage error (PE) and absolute percentage error (APE) are defined as:

$$PE = \frac{(F - A)}{A} 100$$

and

$$APE = \frac{|F - A|}{A} 100$$

where F denotes forecast and A refers to actual population. We therefore assume that the actual population number, as given by the ERP, is accurate. MedAPE is our preferred measure of typical error over mean absolute percentage error (MAPE) because APE distributions from population forecasts generally include a long tail of large errors which influence the mean. It is also used in many studies of population forecast accuracy (e.g., Wilson et al., 2018). To enable comparison of our results with those in other papers using only mean percentage error measures, we also present tables of MAPEs and mean algebraic percentage errors (MALPE) with the supplementary information, where MALPE is the mean of the signed percentage errors.

We also report the percentage of bad forecasts, a measure defined specifically for this study. It is the percentage of small area population forecasts which exceed 10% APE after 5 years and 20% after 10 years. Whilst achieving low values for average error measures is desirable, it is also best to avoid having forecasts with large errors given that important local planning decisions are made using this data. Although 10% and 20% are somewhat arbitrary, they exceed the errors defined by most population forecast users consulted by Wilson and Shalley (2019) on what constitutes an acceptable level of error.

2.5. Geographical classifications

We present results for SA2 areas categorized by remoteness in Australia (ABS, 2013b) and a modification of the urban-rural indicator for New Zealand (Statistics New Zealand, 2020b). This is to determine if some forecasting methods work well in some parts of the settlement hierarchy better than others. In Australia, areas are classified as major cities, inner regional, outer regional, remote, and very remote. We created a corresponding classification system for New Zealand using the urban-rural geography classification (Statistics New Zealand, 2020c) and their concordances to SA2s (Statistics New Zealand, 2020d), such that each area was classified as a major urban area, large urban area, medium urban area, small urban area and rural. These area classifications are a minor modification of the urban-rural indicator (Statistics New Zealand, 2020b). The classification of New Zealand SA2 areas and the method for creating them, and the classification of Australian SA2 areas, are provided in the supplementary material.

3. Results

3.1. Overall error

Table 2 presents a summary of the SA2 area population forecast errors for Australia. MedAPE is shown in the top half of the table, whilst percentage bad forecasts are reported beneath it. Results from the unconstrained forecasts are shown in the upper panel whilst those from the constrained forecasts are shown below them. The best result in each set of forecasts is printed in bold. M4 method results that beat one benchmark model are highlighted yellow, and those that beat both benchmarks are highlighted green.

For Australia, the MedAPEs from all methods were all well under the ‘bad forecasts’ cut-off of 10% after 5 years and 20% after 10 years. Most of the M4 methods did as well as, or better than, the benchmark models. Constraining generally improved forecast accuracy for the 2006-based forecasts, whilst decreasing the accuracy of the 2011-based forecasts. The exception is the Smyl method which was improved for both jump-off years. The addition of a remoteness category as a feature generally improved accuracy for the Smyl method for unconstrained

Table 2
Evaluation of SA2 area total population forecast errors after 5 and 10 years, Australia.

	Benchmark		M4 competition methods										
	LIN/ EXP	CSP- VSG	Smyl	Smyl (r - c)	M-M	Pawl	J & P	F & L	P & S	Shaub	L & K	Door	Pedr
MedAPE (%)													
Unconstrained													
After 5 years													
2006-based	4.2	**3.4**	4.2	3.7	4.1	4.0	4.1	4.1	4.1	4.1	4.6	4.5	4.6
2011-based	**3.3**	4.3	4.9	4.6	**3.3**	**3.3**	3.4	**3.3**	3.4	3.4	3.7	3.6	3.7
After 10 years													
2006-based	7.3	7.1	7.5	**6.3**	7.0	7.0	7.2	7.0	7.3	7.1	7.7	7.6	7.7
Constrained													
After 5 years													
2006-based	3.6	3.4	**3.2**	3.4	3.4	3.3	3.4	3.3	3.4	3.4	3.5	3.7	3.7
2011-based	3.6	4.3	4.1	3.9	**3.5**	3.6	**3.5**	3.7	3.7	3.9	4.4	3.7	3.8
After 10 years													
2006-based	6.5	7.1	6.1	6.5	6.2	6.1	6.1	**6.0**	6.4	6.1	6.7	6.6	6.6
Percentage Bad Forecasts													
Unconstrained													
After 5 years													
2006-based	15.9	**11.9**	14.5	13.9	13.4	13.0	14.6	14.0	13.5	13.4	17.1	16.7	16.4
2011-based	11.9	14.4	15.3	14.2	11.8	11.9	12.0	11.2	**11.1**	11.2	14.4	13.7	14.1
After 10 years													
2006-based	12.6	11.1	11.8	10.4	11.3	**10.8**	11.6	10.9	11.5	11.0	12.8	13.9	13.4
Constrained													
After 5 years													
2006-based	14.8	11.9	11.7	12.4	11.9	**11.4**	12.1	12.3	13.0	**11.4**	12.9	14.7	14.4
2011-based	14.4	14.4	12.8	11.6	13.4	12.8	13.5	**12.1**	12.6	12.8	15.3	14.5	14.7
After 10 years													
2006-based	12.2	11.1	10.3	10.4	10.7	**10.1**	10.5	10.4	11.5	10.8	11.3	13.7	13.1
Total error													
Unconstrained	55.1	52.2	58.2	53.1	50.9	**50.0**	52.9	50.5	50.8	50.2	60.4	60.1	59.7
Constrained	55.1	52.2	48.2	48.2	49.1	**47.4**	49.1	47.8	50.6	48.4	54.2	56.9	56.1

Notes. Boxes are shaded yellow if the result improves on one benchmark, and green if it improves on both. The best result in each row is bolded. Smyl (r-c) = Smyl method with remoteness categories, M-M = Montero-Manso et al.; Pawl = Pawlikowski & Chorowska.; J & P = Jaganathan & Prakash; F & L = Fiorucci & Louzada; P & S = Petropoulos & Svetunkov; L & K = Legaki & Koutsouri; Door = Doornik et al.; Pedr = Pedregal et al. The benchmark CSP-VSG model is automatically constrained to the national population forecast, so there is no separate unconstrained version.

forecasts. However, constraining improved results for the Smyl method more than for the Smyl (r-c) method, but this benefit was reduced or removed with constraining.

The evaluation in terms of percentage bad forecasts paints a similar picture. Although MedAPEs were generally low, the error distributions are such that 10%–17% of all SA2 areas have APE exceeding the bad forecasts threshold.

To summarize overall error patterns for practitioners, we calculated a 'total error' measure at the bottom of Table 2 which is simply the sum of all values of MedAPE and percentage bad forecasts from each method's unconstrained and constrained forecasts as shown in the table. The Smyl method benefitted most from constraining. Whilst the addition of a remoteness categorical variable improved forecast accuracy in the unconstrained condition, this benefit disappeared with constraining. Five of the M4 methods performed marginally better than the CSP-VSG benchmark in both the unconstrained and constrained forecasts. Pawlikowski & Chorowska's method gives the lowest errors (by the tiniest of margins).

Table 3 presents equivalent results for New Zealand. Most of the M4 methods performed as well as the benchmarks, although only the Smyl method was able to improve upon the CSP-VSG forecasts, and only for the 2015-based forecasts. In terms of percentage bad forecasts, most methods produced forecasts with 10%–15% of areas exceeding the bad forecasts threshold. Most forecasts benefitted from constraining, although the Smyl method did not. The addition of a remoteness feature did not improve forecast accuracy for the Smyl method. The CSP-VSG benchmark method performed well overall.

3.2. Error by area characteristics

We also examined forecast errors by population size categories, remoteness area (Australia) and urban-rural

Table 3
Evaluation of SA2 area total population forecast errors after 5 and 10 years, New Zealand.

	Benchmark		M4 competition methods										
	LIN/ EXP	CSP- VSG	Smyl	Smyl (r-c)	M-M	Pawl	J & P	F & L	P & S	Shaub	L & K	Door	Pedr
MedAPE (%)													
Unconstrained													
After 5 years													
2010-based	3.9	**3.0**	3.4	4.3	3.9	3.8	3.8	3.8	3.8	3.7	3.9	3.9	3.8
2015-based	4.4	3.6	**3.3**	3.4	4.3	4.0	4.2	4.3	4.5	4.6	5.4	4.8	5.2
After 10 years													
2010-based	8.4	**6.5**	7.0	9.9	8.2	8.3	8.5	8.2	8.2	8.0	8.9	8.7	8.4
Constrained													
After 5 years													
2010-based	4.1	**3.0**	3.5	3.7	3.5	3.5	3.5	3.5	3.7	3.4	3.3	4.0	3.5
2015-based	4.2	3.6	**3.3**	3.3	4.3	4.0	4.7	4.1	4.6	4.4	4.0	4.6	4.7
After 10 years													
2010-based	8.8	**6.5**	7.5	8.0	7.7	8.0	8.1	7.9	8.2	7.7	7.6	9.9	8.4
Percentage Bad Forecasts													
Unconstrained													
After 5 years													
2010-based	14.8	**10.8**	12.3	13.7	11.9	11.4	12.4	11.1	12.0	11.0	12.1	13.5	12.9
2015-based	18.0	14.6	11.6	**11.4**	16.6	15.8	16.7	17.0	17.3	17.4	22.4	19.1	21.4
After 10 years													
2010-based	14.3	**9.8**	11.0	13.1	11.8	12.6	12.9	12.5	12.4	12.3	12.0	14.2	13.0
Constrained													
After 5 years													
2010-based	15.1	**10.8**	12.4	13.6	11.0	11.4	11.9	11.1	11.8	**10.8**	11.8	13.6	12.5
2015-based	17.3	14.6	**13.0**	12.9	16.4	15.6	18.3	15.7	17.7	16.7	16.3	18.4	19.8
After 10 years													
2010-based	15.3	**9.8**	11.4	11.1	10.8	12.3	12.5	11.9	12.4	11.2	11.2	15.7	13.0
Total error													
Unconstrained	63.8	**48.2**	48.6	55.8	56.7	55.9	58.5	56.9	58.2	57.0	64.7	64.2	64.7
Constrained	64.8	**48.2**	51.1	52.6	53.7	54.8	59.0	54.2	58.4	54.2	54.2	66.2	61.9

Notes. See Table 2.

categories (New Zealand), and recent population growth rates. Previous research has shown that errors are often higher for smaller populations, more remote areas, and areas with very high or very low growth rates in the base period (Tayman, 2011; Wilson & Rowe, 2011). The tables for these disaggregated results are large and are contained in the supplementary material accompanying the paper.

Forecast errors for SA2s in Australia for four jump-off year population size categories (populations of 0–4,999, 5000–9000, 10,000–14,999, and 15,000+) are presented in Table S1a. Although many previous studies have found a negative association between population size and error, this relationship is only evident here for the benchmark CSP-VSG model. The top-performing methods differed by jump-off year and population size. Constraining was more helpful for larger areas than smaller areas and for the 2006-based forecasts more than the 2011-based forecasts.

Forecast errors for four population size categories for New Zealand SA2 areas (0–999, 1000–1999, 2000–2000 and 3000+) are shown in Table S1b. There is a clearer association between error and population size, possibly because the populations are smaller. Previous research has shown error varies most by population size at the very lowest end of the population size range (Wilson et al., 2018). The results are similar to the findings from Table 2: the CSP-VSG benchmark model performs a little better than others in the 2010-based forecasts whilst Smyl does best in the 2015-based forecasts. These findings are consistent across the constrained and unconstrained versions of the forecasts and across population size categories. Constraining appears to have less impact for New Zealand than for Australia.

Forecast errors by remoteness area categories in Australia are shown in Table S2a. Most methods forecast the inner regional and outer regional areas the best and very remote the worst. There is more variation in errors between methods relative to the population size categories. In the unconstrained forecasts, the benchmark CSP-VSG model did quite well for major cities in the 2006-based forecasts, but not in the 2011-based forecasts. Legaki & Koutsouri performed well in forecasting 2016 populations for outer regional, remote, and very remote areas. There is no obvious best method for particular types of remoteness area, although the improved 10-year forecast results for the most remote areas by M4 methods are promising. The addition of remoteness categories as an external variable for the Smyl method did not have a consistent effect. For the 2006-based unconstrained forecasts, error decreased for less remote areas and increased for more remote areas; differences were reduced with constraining. The Smyl (r-c)'s forecasts for the remote and very remote areas benefitted from constraining, whilst the Smyl forecasts for these areas worsened after constraining. Constraining tended to be more beneficial for less remote areas.

Forecast errors for urban accessibility areas in New Zealand are shown in Table S2b. Major urban areas and rural areas generally experienced slightly lower errors than other area types. In contrast to Australia, the overall picture is clearer: the benchmark CSP-VSG model was the best for the 2010-based forecasts, whilst Smyl was the top-performing method for the 2015-based forecasts for all area types. Constraining was generally helpful, although there are exceptions (the Smyl method). The addition of the remoteness category feature increased error for unconstrained - but not constrained - forecasts.

The performance of the various methods broken down by growth rate over the decade prior to the jump-off year is shown in Table S3a for Australian SA2 areas. Overall, areas with the highest and lowest growth rates experienced higher errors than those with more moderate growth rates, which is consistent with previous research. M4 methods tended to produce the best results for the areas with the highest and lowest growth rates. Constraining had mixed effects. It was generally more helpful for SA2s with growth <2%, but not for the fastest growing areas. There are exceptions. For example, the Smyl method had a MedAPE of 4.8% for SA2s with growth <1% p.a. in the 2006-based 10-year forecasts; this increased to 7.8% after constraining. For the 2011-based forecasts, constraining was not helpful for most of the methods, except for the Smyl and Smyl(r-c) methods which benefitted from constraining in all the forecasts. Different methods produced the lowest median errors for different growth rate categories. In the unconstrained forecasts, Legaki & Koutsouri gave the lowest errors for areas which had been declining in population, whilst Pawlikowski & Chorowska did well for higher growth rate areas. In the constrained forecasts, Pedregal et al. and the LIN/EXP model produced the lowest errors for declining populations, whilst the Smyl and the Pawlikowski & Chorowska methods were competitive for higher growth areas. However, the variation in errors between the 2006-based and 2011-based forecasts after 5 years, and between the unconstrained and constrained forecasts, suggests that the results are limited in the guidance that they can provide in method selection.

Table S3b reports equivalent results by growth rate category for New Zealand. Again, the Pawlikowski & Chorowska method consistently performed well for higher growth areas. The benchmark CSP-VSG model did well for the 2010-based forecasts, particularly for SA2s with population growth <2% p.a., whilst the Smyl method did well in the 2015-based forecasts. Constraining had less impact on the New Zealand forecasts than on the Australian forecasts; error tended to decrease for lower growth areas and there was minimal impact on high growth areas. The Smyl and Smyl (r-c) methods are exceptions. Their 5-year 2006-based forecasts improved by 50% or more with constraining.

3.3. Forecast bias

Our evaluation focuses on forecast accuracy, although we briefly consider forecast bias as measured by MALPE. These are presented beneath MAPEs in supplementary Tables S4a and S4b for Australia and New Zealand, respectively. For Australian forecasts, with few exceptions, MALPEs were generally negative. The benchmark methods tended to yield small MALPEs in both unconstrained and constrained forecasts, although the better benchmark method is different for different forecasts. Most M4 methods tended to have intermediate bias levels, particularly after constraining. The exception was the constrained 2011-based forecasts; several M4 methods do better, of which Doornik et al. have the best result. MALPEs become more positive with constraining, except for the Smyl and Smyl (r-c) 2011-based forecasts. Results are different for the New Zealand MALPEs: 2015-based forecast MALPEs are all negative, whilst most 2010-based forecast MALPEs are positive. The impact of constraining on the MALPE metric varied by jump-off year and method.

3.4. Summed SA2 forecasts

We employed a simple scaling factor to constrain our forecasts. To investigate why constraining was more helpful for some forecasts but not others we consider the ERP for the target years, the national forecasts, and the summed small area forecasts for each of the methods evaluated. The results are presented in Supplementary Table S5. The national forecasts for Australia and New Zealand all under-predicted the ERP, except for the Australian 2011-based forecast for 2016, which over-predicted the national population by 0.71%. The Australian national forecasts proved more accurate than those for New Zealand (errors of −1.3% to 0.7% for Australia and −6.4% to −0.9% for New Zealand). The summed 2011-based SA2 forecasts by the Smyl and Smyl (r-c) methods over-predict the ERP for 2016. Every single other M4 method under-predicted the total ERP regardless of country or jump-off year. Furthermore, almost all summed M4 method forecasts were smaller than the national forecast; exceptions included forecasts by the LIN/EXP, Smyl, Smyl (r-c), Petropoulos & Svetunkov, and Doornik methods, and are mostly for New Zealand.

In the light of our consideration of the summed forecasts, it appears that where the national forecast is closer to the ERP, constraining improves forecast accuracy. Where it was further from the ERP, forecast accuracy was lower. The summed SA2 forecasts are closer to the national forecast for New Zealand than Australia, and for this reason, constraining has a greater impact on the Australian forecasts.

4. Discussion

In this paper, we present an evaluation of small area demographic forecasts for Australia and New Zealand using the top ten M4 methods. We found that:

- The M4 methods performed quite well, generally out-performing at least one of our benchmarks.
- Performance varied between Australian and New Zealand, and for different jump-off years. Most of the top seven M4 methods were able to improve upon both benchmark models for the Australian SA2 forecasts, but only the Smyl method was able to do so for any of the New Zealand forecasts.
- An iterative adaption of the Smyl M4 method was shown to provide good forecasts where the input data time series was short.
- Constraining to national forecasts produced mixed outcomes. The benefit of constraining tended to be greater for larger, less remote areas, and those with smaller growth rates.
- Local constraints need to be incorporated into models for small area population forecasting to prevent runaway growth in individual area forecasts.
- There is no clear winner in our evaluation. Further work is required as all methods produced unacceptably high errors for a minority of small areas.

We note that both the CSP-VSG and LIN/EXP benchmark methods were found to be strong performers for the forecasting of small area populations in a previous evaluation (Wilson, 2015). Therefore, the result that most of the M4 methods perform as well as at least one benchmark is promising because practitioners will not know in advance which method will perform better. Top methods varied for the different countries and jump-off years, although some effects appear to be generalizable. The CSP-VSG benchmark method tends to do best for areas with intermediate growth rates, whilst M4 methods are better at forecasting areas with low or high growth. Despite overall MALPEs sometimes being positive, summed small area forecasts using M4 methods generally underestimate both the target year's ERP and the national forecast for the target year. The Pawlikowski & Chorowska method consistently performs well for areas with high growth rates for both Australian and New Zealand SA2s. The Smyl method regularly, but not consistently, outperforms both benchmark methods.

The CSP-VSG method tends to be the better performing benchmark method, particularly for New Zealand. The CSP-VSG method is a combination of the CSP and VSG methods, with the latter being an adjusted LIN/EXP method which produces forecasts similar to it. The LIN/EXP method did not do particularly well for the New Zealand dataset; the improved performance of the CSP-VSG method is due to its CSP component. The CSP method considers that an area's share of the population is fixed over time (i.e., the national population growth rate applies to each small area) and uses the national forecast to produce the forecasts, suggesting that small area population growth in New Zealand is less variable than in Australia.

The Pawlikowski & Chorowska method is the top overall performer for the Australian SA2 forecasts and performed well for the New Zealand forecasts. It did not do particularly well for areas with smaller populations or those that were more remote. However, it consistently outperformed other methods for small areas with growth rates above 2%, although constraining tended to increase error. The good performance occurs even though the model pool and weights were chosen for the M4 time series, suggesting that the method has good transferability. Therefore, it may be a good method to recommend for forecasting small areas with historic high growth rates.

The Smyl method performs well, even where an iterative adaption of the method is applied to overcome issues linked with the short time series. There is no generalizable improvement with the addition of a remoteness category feature to the Smyl method. Smyl (r-c) did better than the regular Smyl for the unconstrained Australian forecasts, although this benefit disappeared with constraining. For the 2006-based forecasts, the addition of the remoteness category improved the results for less remote areas, whilst worsening remote and very remote SA2 results. Conversely, for the 2011-based forecasts, results for major cities and very remote areas worsened, while those in other remoteness categories improved. This pattern was not seen with New Zealand SA2s. The Smyl (r-c) method performed worse than the Smyl method and did not improve forecast accuracy for any of the remoteness areas, however, the differences were reduced with constraining. Perhaps given that New Zealand is significantly smaller in size than Australia, remoteness does not have the same amount of bearing on small area population growth. Small area population growth rates are less variable (as supported by good performance of the CSP model), and the addition of external variables merely limits the amount of potentially useful cross-learning. However, we also note that the presence of potential confounding influences the remoteness area classifications, distributions vary between countries, and our classification of New Zealand areas involves a minor modification to the official urban-rural indicator to resolve issues where areas are given multiple classifications.

Constraining had the greatest impact on the Smyl method. The total error metric shows that constraining improves the overall accuracy of the Smyl method by 17% for Australia but reduces accuracy by 5% for the New Zealand forecasts. Given its strong performance in our evaluation, we provide several recommendations for further adaptions to support work with small area datasets. First, the original Smyl method is trained with an output window which is the same width as the forecast horizon. Unfortunately, small area datasets are rarely long enough to support this, and an iterative approach is required. An iterative approach also allows a constraining method to be incorporated at each step. Whilst the addition of our remoteness category feature did not produce consistent results, it may be because remoteness is the wrong feature to include. A future approach could involve clustering the time series by their characteristics and running different global models for each cluster such as in Bandara, Bergmeir, and Smyl (2020), or using the cluster identities as the external variables (Bandara, Shi, Bergmeir, Hewamalage, Tran, & Seaman, 2019). The advantage of this approach is that it does not require further data outside of the population counts. This is important because the available features differ between small areas around the world. Furthermore, models for forecasting population totals can use more recent data as reporting of other details generally lags in time.

The Montero-Manso et al. (2020) method was the top performer for forecasting demographic data in the M4 competition and performed well on our small area data. However, it did not stand out amongst the other top seven methods. The Montero-Manso and Shaub methods both produced significant errors for the Prestons SA2 area for the 2015-based forecast. There is a sudden increase in population from 2014 (n = 300) to 2015 (n = 1,100) due to the construction of a new housing development (Mitchell, 2013). It may be useful for methods used for demographic forecasts to include local constraints for each series to prevent this.

Runaway growth for the Prestons area was readily identifiable, and for this reason, those results are removed from our evaluation. Forecasts for the New Zealand area Otakaro-Avon River Corridor are also highly erroneous, producing errors >8000% for all the methods in the 2010-based forecasts for 2015, and >10,000% for all the methods in 2010-based for 2020. Its population decreased from 10,950 in 2010 to 130 in 2015 and 100 in 2020 because this area was severely impacted by the Canterbury earthquakes of 2010 and 2011 (Department of the Prime Minister and Cabinet, 2019). The standard approach would be to ignore this result as an unpredictable outlier. However, if forecasts are being produced over many areas, over 10, 20, or more years, then at least some will be impacted by natural disasters and other extreme events over this time. They also have knock-on effects. For example, the fast-paced housing development in Prestons was approved through emergency legislation to rehouse those who were displaced by the Canterbury earthquakes (Aurecon, 2018). In the post-COVID-19 world, there needs to be greater consideration of such events. Whilst it may not be possible to predict them, perhaps one can attempt to build in a certain level of robustness for them. For example, if hierarchical forecasts are being created to support the construction of critical infrastructure, then enough capacity should be available across the system at higher geographies to support issues at the small area level.

In this paper, we focused on the forecasting of small area populations. These forecasts can be used as is, although in practice they are often constrained to national or regional forecasts to provide a set of forecasts fully consistent across geographical scales (e.g., South Australian Department of Planning, Transport and Infrastructure, 2020). In our study, we found that the national forecast tended to be more accurate than summed small area forecasts, and that constraining using simple factor scaling tended to improve forecast accuracy overall. This is in line with other literature; Panagiotelis, Athanasopoulos, Gamakumara, & Hyndman, 2021, provide a geometric interpretation of reconciliation methods showing that they improve forecast accuracy. However, this was not always the case with our forecasts, and the constraining and forecasting methods interacted. For example, the difference between the Smyl and Smyl (r-c) forecasts are decreased with constraining. The impact of constraining varies with area size, remoteness, and method used. Further research is required, and the exploration of more sophisticated forecast reconciliation methods are warranted, examples include Hyndman, Ahmed, Athanasopoulos, and Shang (2011) and Wickramasuriya, Athanasopoulos, and Hyndman (2019).

We focused only on 5- and 10-year forecast horizons. Longer horizons would not have enabled the use of a retrospective approach to determine forecast accuracy. We note that 5- and 10-year forecasts are desired horizons for many users. A past survey of the users of Australian population projections found that only 36% used forecasts exclusively greater than 5 years (Diamond, Tesfaghiorghis, & Joshi, 1990). Forecasts for small areas often have MAPEs greater than 10% at 10 years (Wilson et al., 2018), and MAPEs generally increase by 10%–14% per decade (Rayer & Smith, 2010). For areas with population < 2500, an error of 40% is common for 20-year forecast horizons (Rayer & Smith, 2010). These high errors are consistent with the MAPEs in our study, which are presented in the supplementary information. Small area population forecasts beyond 10 years are increasingly unreliable.

Practitioners in various fields have previously commented on the M4 forecasting competition, including Google data scientists (Fry & Brundage, 2020). They note that (1) time intervals are changing, (2) time series are often hierarchical, (3) forecasts need more information than just time series, (4) prediction intervals are important, and (5) one size does not fit all. The five key points described by Fry and Brundage (2020) are also largely applicable to demographic forecasting, although there are some context-specific differences. As discussed in this paper, forecasts often also need to be hierarchical. The recent M5 competition focused on hierarchical forecast methods (Makridakis, Spiliotis, & Assimakopoulos, 2020b), and these methods would be worth investigating for small area forecasting. Prediction intervals are also important, as are simple methods of communicating the meaning of these intervals to users. The concept of a forecast shelf-life, beyond which forecast uncertainty becomes too great for reliable predictions, has been proposed (Simpson, Wilson, & Shalley, 2018; Wilson, 2018). Fry and Brundage (2020) describe the usage of higher frequency data. Many demographers often continue to rely on census data collected every 5 or 10 years. There is a movement to support the provision of accurate yearly demographic time series. Goal 17.18 of the UN's sustainability goals involves supporting developing countries in the collection of demographic data disaggregated by geographic location (United Nations, 2018). Sparse data is also a significant issue in small area demography, as is erroneous input data. The ERP used as input data are estimates, not precise counts of population. The accuracy of these data is variable and tends to be lower for smaller and more remote areas. For this reason, Bayesian methods are beginning to become more widely used in small area demography (Wilson, Grossman, Alexander, Rees, & Temple, 2021).

Forecasting research is characterized by the development of sophisticated, often computationally expensive methods, and developed with the goal of improving forecast accuracy. These methods are often beyond the reach of many practitioners. In demography, issues other than accuracy are important in selecting population forecasting methods (Smith et al., 2013; Wilson, 2011). Factors that need to be considered include the amount of input data needed, the computing software and hardware requirements, computer processing time, staff expertise required, person-hours needed to prepare the forecasts, the budget available, and transparency of methods. Whilst it is true that these issues are also relevant in other fields, resources are often more limited in the government sector than in the business sector. The M4 methods all require considerably more data preparation, staff time, computing equipment and time, and overall effort to produce than the simple benchmark models. However, if methods can be shown to consistently improve upon existing methods, they have a greater chance of being implemented.

Recent years have seen important advances in time series forecasting methods - as demonstrated by the improvements in forecast accuracy over the course of the M Competitions (Makridakis, Hyndman, & Petropoulos, 2020). With time, it is expected that some of the new methods should be capable of producing more accurate demographic forecasts. Forecasting researchers need to engage with practitioners to determine selection criteria for new forecasting models. In addition, methods need to be tested on multiple datasets and multiple jump-off years. Here, we only considered Australian and New Zealand small areas, and over a relatively short time period. A more comprehensive evaluation would require multiple small area demographic datasets. To fulfil this need, we are working to create a repository of historical small area demographic estimates from many countries to facilitate the testing of new methods, and we invite demographers and forecasting researchers to contribute and use these datasets (https://demographic-datasets-network.github.io/).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of data

SA2 area population forecasts, SA2 area Estimated Resident Populations, and Remoteness and Urban-Rural classifications are available in the supplementary material accompanying this paper. The code for the methods is available in the M4 Competition Github, https://github.com/Mcompetitions/M4-methods.

Acknowledgments

We would like to thank the authors of the M4 methods who assisted us in the implementation of their models. Dr Kasun Bandara kindly provided helpful comments on an earlier draft of the paper. All errors and omissions, however, remain the authors' responsibility

Funding source

This research was supported by the Australian Research Council (Discovery Project DP200101480).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2021.09.005.

References

Aurecon (2018). Prestons subdivision, christchurch, New Zealand. https://www.aurecongroup.com/projects/property/prestons-subdivision-christchurch. (Accessed 21 June 2021).

Australian Bureau of Statistics (2008). *TABLE B9. Population projections, By age and sex, Australia - Series B. Australian Bureau of Statistics repository*, https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3222.02006%20to%202101?OpenDocument.

Australian Bureau of Statistics (2013a). *TABLE B9. Population projections, By age and sex, Australia - Series B, Australian Bureau of Statistics repository*, https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3222.02012%20(base)%20to%202101?OpenDocument.

Australian Bureau of Statistics (2013b). *1270.0.55.005 - Australian Statistical Geography Standard (ASGS): Volume 5 - Remoteness structure, 2011*. Australian Bureau of Statistics repository, https://www.abs.gov.au/AUSSTATS/abs@.nsf/allprimarymainfeatures/17A7A350F48DE42ACA258251000C8CA0?opendocument.

Australian Bureau of Statistics (2017). *ERP by SA2 and above (ASGS 2011), 1991 to 2016, Australian Bureau of Statistics ABS.Stat Beta*, http://stat.data.abs.gov.au/Index.aspx?DataSetCode=ABS_ANNUAL_ERP_ASGS.

Baker, J., Alcántara, A., Ruan, X., Watkins, K., & Vasan, S. (2014). Spatial weighting improves accuracy in small-area demographic forecasts of urban census tract populations. *Journal of Population Research*, *31*, 345–359.

Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, *140*, Article 112896.

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019). Sales demand forecast in E-commerce using a long short-term memory neural network methodology. In T. Gedeon, K. Wong, & M. Lee (Eds.), *Lecture notes in computer science*: *Vol. 11955*, *Neural information processing. ICONIP 2019* (pp. 462–474). Cham: Springer.

Business Forecast Systems, Inc. (2018). Forecast Pro TRAC. Version:5.1. https://www.forecastpro.com/. (Accessed 21 Fenruary 2021).

Chen, Y., Li, X., Huang, K., Luo, M., & Gao, M. (2020). High-resolution gridded population projections for China under the shared socioeconomic pathways. *Earth's Future*, *8*, Article e2020EF001491.

Chi, G., & Voss, P. R. (2011). Small-area population forecasting: Borrowing strength across space and time. *Population, Space and Place*, *17*, 505–520.

Department of the Prime Minister and Cabinet (2019). *Ōtākaro Avon River Corridor Regeneration Plan*. New Zealand Government, https://dpmc.govt.nz/our-programmes/greater-christchurch-recovery-and-regeneration/recovery-and-regeneration-plans/otakaro-avon-river-corridor-regeneration-plan. (Accessed 21 June 2021).

Diamond, I., Tesfaghiorghis, H., & Joshi, H. (1990). The uses and users of population projections in Australia. *Journal of the Australian Population Association*, *7*(2), 151–170.

Doornik, J. A. (2013). *Object-oriented matrix programming using Ox* (7th ed.). London: Timberlake Consultants Press.

Doornik, J. A., Castle, J. L., & Hendry, D. F. (2020). Card forecasts for M4. *International Journal of Forecasting*, *36*, 129–134.

Doornik, J. A., & Ooms, M. (2006). *Introduction to ox: An object-oriented matrix language*. London: Timberlake Consultants Press.

Fiorucci, J. A., & Louzada, F. (2020). Groec: combination method via generalized rolling origin evaluation. *International Journal of Forecasting*, *36*, 105–109.

Fiorucci, J., Louzada, F., & Yiqi, B. (2016). *Forectheta: forecasting time series by theta models*. R package version, 2.

Fry, C., & Brundage, M. (2020). The M4 forecasting competition - A practitioner's view. *International Journal of Forecasting*, *36*, 156–160. http://dx.doi.org/10.1016/j.ijforecast.2019.02.013.

Hachadoorian, L., Gaffin, S. R., & Engelman, R. (2011). Projecting a gridded population of the world using ratio methods of trend extrapolation. In R. P. Cincotta, & L. J. Gorenflo (Eds.), *Human population: its influences on biological diversity* (pp. 13–25). Berlin: Springer-Verlag.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*(9), 2579–2589.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., et al. (2018). *Forecast: Forecasting functions for time series and linear models. R package version 8.4*. https://CRAN.R-project.org/package=forecast.

Jaganathan, S., & Prakash, P. (2020). A combination-based forecasting method for the M4-competition. *International Journal of Forecasting*, *36*, 98–104.

Kuang, K., Kong, Q., & Napolitano, F. (2019). pbmcapply: Tracking the Progress of Mc*pply with Progress Bar. https://cran.r-project.org/web/packages/pbmcapply/index.html. (Accessed 2 February 2021).

Legaki, N.-Z., & Koutsouri, A. (2018). Method description. https://github.com/Mcompetitions/M4-methods/blob/master/260%20-%20KaterinaKou/M4-Method-Description_NZL.pdf. (Accessed 2 February 2021).

Makridakis, S., Hyndman, R. J., & Petropoulos, F. (2020). Forecasting in social settings: The state of the art. *International Journal of Forecasting*, *36*(1), 15–28.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020a). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*, 54–74.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020b). The M5 accuracy competition: Results, findings and conclusions. *International Journal of Forecasting*.

Mitchell, I. (2013). *Greater Chirstchurch housing market assessment*: *Research report*, https://www.greaterchristchurch.org.nz/assets/Documents/greaterchristchurch/Projects/Greater-Christchurch-Housing-Market-Assessment.pdf. (Accessed 21 June 2021).

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*, 86–92.

Neubig, G., Dyer, C., Goldberg, Y., Matthews, A., Ammar, W., Anastasopoulos, A., et al. (2017). Dynet: The dynamic neural network toolkit. arXiv preprint arXiv:1701.03980.

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., & Hyndman, R. J. (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting*, *37*(1), 343–359.

Pawlikowski, M., & Chorowska, A. (2020). Weighted ensemble of statistical models. *International Journal of Forecasting*, *36*, 93–97.

Pedregal, D. J., Trapero, J. R., Villegas, M. A., & Madrigal, J. J. (2018). Method description. https://github.com/Mcompetitions/M4-methods/blob/master/039%20-%20djpt999/M4-Method-Predilab.pdf. (Accessed 1 February 2021).

Petropoulos, F., & Makridakis, S. (2020). The M4 competition: Bigger. Stronger. Better. *International Journal of Forecasting*, *36*(1), 3–6.

Petropoulos, F., & Svetunkov, I. (2020). A simple combination of univariate models. *International Journal of Forecasting*, *36*(1), 110–115.

Rayer, S., & Smith, S. K. (2010). Factors affecting the accuracy of subcounty population forecasts. *Journal of Planning Education Research*, *30*(2), 147–161.

Reinhold, M., & Thomsen, S. L. (2015). Subnational population projections by age: An evaluation of combined forecast techniques. *Population Research and Policy Review*, *34*, 593–613.

Riiman, V., Wilson, A., Milewicz, R., & Pirkelbauer, P. (2019). Comparing artificial neural network and cohort-component models for population forecasts. *Population Review*, *58*.

Shaub, D. (2020). Fast and accurate yearly time series forecasting with forecast combinations. *International Journal of Forecasting*, *36*, 116–120.

Shaub, D., & Ellis, P. (2018). *ForecastHybrid: Convenient functions for ensemble time series forecasts*. R package version 3.0.14.

Simpson, L., Wilson, T., & Shalley, F. (2018). *The shelf life of subnational population forecasts, from Australia To England*: *Working paper*, Northern Institute, Charles Darwin University.

Smith, S. K., Tayman, J., & Swanson, D. A. (2013). *A practitioner's guide to state and local population projections*. Springer.

Smyl, S. (2017). *ES-RNN-E: Exponential Smoothing Recurrent Neural Network hybrid, Ensemble of specialists. Point forecast [Source code]*. https://github.com/Mcompetitions/M4-methods/tree/master/118%20-%20slaweks17.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*, 75–85.

South Australian Department of Planning, Transport and Infrastructure (2020). *Local Area (SA2 and LGA) Population Projections for South Australia, 2016 to 2036*. Government of South Australia, https://plan.sa.gov.au/__data/assets/pdf_file/0010/822727/Local_Area_SA2_and_LGA_Population_Projections_for_South_Australia_2016_to_2036.pdf.

Statistics New Zealand (2009). *National Population Projections: 2009(base)-2061 - Tables. Statistics New Zealand repository*, https://catalogue.data.govt.nz/dataset/national-population-projections.

Statistics New Zealand (2014). *National Population Projections: 2014(base)-2068 - table. Statistics New Zealand repository*, https://catalogue.data.govt.nz/dataset/national-population-projections.

Statistics New Zealand (2020a). *Subnational population estimates (RC, SA2), by age and sex, at 30 1996-2020 (2020 boundaries), Statistics New Zealand NZ.Stat*, http://nzdotstat.stats.govt.nz/wbos/Index.aspx?DataSetCode=TABLECODE7979#.

Statistics New Zealand (2020b). *Urban accessibility - methodology and classification*, https://www.stats.govt.nz/methods/urban-accessibility-methodology-and-classification. (Accessed 1 February 2021).

Statistics New Zealand (2020c). *Urban rural 2020 (generalized), statistics new zealand datafinder*, https://datafinder.stats.govt.nz/layer/104269-urban-rural-2020-generalised/.

Statistics New Zealand (2020d). *Statistical Area 2 2020 to Urban Rural 2020 V1.0.0, Statistics New Zealand Ariā*, http://aria.stats.govt.nz/aria/?_ga=2.119585928.1895754239.1575164187-283711881.1571107602#ConcordanceView:uri=http://stats.govt.nz/cms/ConcordanceVersion/EiXgaLjmM2XW5XKz.

Tayman, J. (2011). Assessing uncertainty in small area forecasts: State of the practice and implementation strategy. *Population Research and Policy Review*, *30*, 781–800.

United Nations (2018). *Goal 17: Strengthen the means of implementation and revitalize the Global Partnership for Sustainable Development*, https://unstats.un.org/sdgs/metadata/files/Metadata-17-18-02.pdf. (Accessed 21 June 2021).

Weber, H. (2020). How well can the migration component of regional population change be predicted? A machine learning approach applied to german municipalities. *Comparative Population Studies*, *45*.

Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, *114*(526), 804–819.

Wilson, T. (2011). *A review of sub-regional population projection methods*: *Report to the office of economic and statistical research. Queensland centre for population research, school of geography, planning and environmental management*, The University of Queensland, Brisbane.

Wilson, T. (2014). Simplifying local area population and household projections with POPART. In N. Hoque, & L. Potter (Eds.), *Emerging Techniques in Applied Demography* (pp. 25–38). Dordrecht: Springer.

Wilson, T. (2015). New evaluations of simple models for small area population forecasts. *Population, Space and Place*, *21*, 335–353.

Wilson, T. (2016). Evaluation of alternative cohort-component models for local area population forecasts. *Population Research and Policy Review*, *35*, 241–261.

Wilson, T. (2018). *Communicating Population Forecast Uncertainty Using Perishable Food Terminology (Research Brief RB03/2018)*. Northern Institute, Charles Darwin University, http://www.cdu.edu.au/sites/default/files/research-brief-2018-03_0.pdf.

Wilson, T., Brokensha, H., Rowe, F., & Simpson, L. (2018). Insights from the evaluation of past local area population forecasts. *Population Research and Policy Review*, *37*, 137–155.

Wilson, T., Grossman, I., Alexander, M., Rees, P., & Temple, J. (2021). Methods for small area population forecasts: state-of-the-art and research needs. *Population Research and Policy Review*, http://dx.doi.org/10.1007/s11113-021-09671-6.

Wilson, T., & Rowe, F. (2011). The forecast accuracy of local government area population projections: a case study of queensland. *Australasian Journal of Regional Studies*, *17*, 204–243.

Wilson, T., & Shalley, F. (2019). Subnational population forecasts: do users want to know about uncertainty? *Demographic Research*, *41*(13), 367–392. http://dx.doi.org/10.4054/DemRes.2019.41.13.