Author/s:
Grossman, I;Bandara, K;Wilson, T;Kirley, M

Title:
Can machine learning improve small area population forecasts? A forecast combination approach

Date:
2022-04-19

Citation:
Grossman, I., Bandara, K., Wilson, T. & Kirley, M. (2022). Can machine learning improve small area population forecasts? A forecast combination approach. Computers, Environment and Urban Systems, 95, https://doi.org/10.1016/j.compenvurbsys.2022.101806.

Persistent Link:
https://hdl.handle.net/11343/332474

Can machine learning improve small area population forecasts? A forecast combination approach

Irina Grossman[a,*], Kasun Bandara[b], Tom Wilson[a], Michael Kirley[b]

[a]*Melbourne School of Population and Global Health, University of Melbourne, Australia*
[b]*School of Computing and Information Systems, Melbourne Centre for Data Science, University of Melbourne, Australia*

[*]Corresponding Author Name: Irina Grossman, Affiliation: Melbourne School of Population and Global Health, The University of Melbourne, Postal Address: 207 Bouverie St, Melbourne, Vic 3010, Australia, The University of Melbourne, Victoria 3052, Australia, E-mail address: Irina.Grossman@unimelb.edu.au

Can machine learning improve small area population forecasts? A forecast combination approach

Abstract

Generating accurate small area population forecasts is vital for governments and businesses as it provides better grounds for decision making and strategic planning of future demand for services and infrastructure. Small area population forecasting faces numerous challenges, including complex underlying demographic processes, data sparsity, and short time series due to changing geographic boundaries. In this paper, we propose a novel framework for small area forecasting which combines proven demographic forecasting methods, an exponential smoothing based algorithm, and a machine learning based forecasting technique. The proposed forecasting combination contains four base models commonly used in demographic forecasting, a univariate forecasting model specifically suitable for forecasting yearly data, and a globally trained Light Gradient Boosting Model (LGBM) that exploits the similarities between a collection of population time series. In this study, three forecast combination techniques are investigated to weight the forecasts generated by these base models. We empirically evaluate our method, by preparing small area population forecasts for Australia and New Zealand. The proposed framework is able to achieve competitive results in terms of forecasting accuracy. Moreover, we show that the inclusion of the LGBM model always improves the accuracy of combination models on both datasets, relative to combination models which only include the demographic models. In particular, the results indicate that the proposed combination framework decreases the prevalence of relatively poor forecasts, while improving the reliability of small area population forecasts.

Keywords: Population Forecasts, Nowcasting, Small Area Population Forecasting, Forecast Combinations, Light Gradient Boosting Model

1. Introduction

Generating accurate and reliable projections of future population is essential for the public and private sectors to proactively allocate resources and to meet the demand of various services and infrastructure, such as housing, childcare, education, aged care, health facilities, energy and water demand, transport, and electoral redistricting (Shafizadeh-Moghadam, 2019; Hasegawa et al., 2019; Badmos et al., 2019). A successful resource planning strategy enables high-quality services to be delivered to the public in a timely manner. Nevertheless, current practices available for generating population forecasts are susceptible to producing highly inaccurate numbers. This can largely be attributed to the challenging nature of population forecasting: the underlying demographic processes are complex, the data may be sparse or unavailable at the small area level, and time series are often short due to frequently changing geographic boundaries. Although several demographic models have been proposed to forecast better under these circumstances, there is significant potential to improve the accuracy in small area forecasting. A recent review by Wilson et al. (2021a) identifies several areas that require further exploration in population forecasting, including the investigation of forecast combination methods and the inclusion of machine learning based forecasting methods.

Practitioners generating combination forecasts are expected to solve two problems. First, they need to decide which individual models to include in the forecast combination, i.e., the base models. Secondly, they must select which combination method to use to assign weights to the constituent forecasts. In the machine learning literature, this procedure is known as “ensembling”. These ensemble models aim to reduce the overall model variance and model bias by aggregating the predictions over multiple models, where the resultant forecasts often yield better accuracy than individual models (Schapire, 1999; Breiman, 2001). A simple mean combination of forecasts is considered as one of the most frequently used ensemble technique in the forecasting literature, often producing competitive results (Timmermann, 2006; Genre et al., 2013). A minor

modification of the simple mean is known as the "trimmed mean", in which the highest and lowest forecasts of the base models are excluded from the simple mean calculation to minimise the impact of extreme values (Rayer et al., 2009; Rayer & Smith, 2010). Instead of weighting each base model forecasts equally, the feature-based forecast combination framework (FFORMA; Montero-Manso et al., 2020) uses a gradient boosted-tree based machine learning model to determine the optimal weight combination for the forecasts produced from multiple base models. In particular, FFORMA is the second-best performing approach in the M4 forecasting competition (Makridakis et al., 2018), and has recently shown competitive results in the population forecasting field (Wilson et al., 2021b).

Although small area population forecasts are used extensively, there is little guidance on the most appropriate models. Share of growth methods, which assume that a small area's share of national or state growth in the base period remains constant throughout the forecast horizon, have produced relatively accurate forecasts for the census tracts in three Florida counties (Smith & Shahidullah, 1995) and Dutch municipalities (Openshaw & Van Der Knaap, 1983). More recently, an evaluation of individual, composite, and simple combination methods show that a composite linear/exponential (LIN/EXP), a simple average of a constant share of population and a variable share of growth model (CSP-VSG), and a modified exponential model (MEX) with floor and ceiling limits, achieve promising results for Australian, New Zealand, English and Welsh small areas (Wilson, 2015).

The literature investigating the application of machine learning algorithms to small area population data is very limited. Recently Riiman et al. (2019) found that the populations of counties in Alabama could be forecast with greater accuracy by training long short-term memory (LSTM) networks separately for each small area time series than by using a traditional cohort-component model. However, time series forecasting research has recently been evolving from traditional univariate models that treat each time series separately to models that are trained across

sets of many related time series. These models are known as global forecasting models (GFM) that build a single model across many time series (Januschowski et al., 2020). GFMs train a unified model that exploits key structures, behaviours, and patterns common within a group of time series (Januschowski et al., 2020; Bandara et al., 2021).

Global models have recently been adapted for small area forecasting, where Grossman et al. (2022) demonstrate that globally trained LSTM networks with automated tuning perform similarly to, or better than the traditional demographic benchmarks for Australian small area populations. The competitive performance of global models has recently been demonstrated in the renowned M4 (Makridakis et al., 2018) and M5 (Makridakis & Spiliotis, 2021) forecasting competitions, and in various domains such as retail (Bandara et al., 2019; Salinas et al., 2020), energy (Bandara et al., 2021; Eshragh et al., 2019), and healthcare forecasting (Bandara et al., 2020a). The winning submission of the M4 competition was Exponential Smoothing-Recurrent Neural Network (ES-RNN; Smyl, 2019), which uses an exponential smoothing model and a recurrent neural network to train across sets of many time series. More recently, Wilson et al. (2021b) show that the ES-RNN model can be a competitive benchmark for small area population forecasting. Meanwhile, the winner of the M5 competition was a globally trained Light Gradient Boosting Model (LGBM; Ke et al., 2017), followed by many other LGBM-based-solutions among the top twenty best performing methods (Makridakis & Spiliotis, 2021). Nonetheless, LGBMs have not been evaluated for small area population forecasting, even though they are becoming increasingly popular machine learning models among forecasting practitioners (Hewamalage et al., 2021a). Moreover, the use of LGBM as a global model enables us to utilise the observations across all population time series, obviating the data availability limitations in a single yearly population time series.

Whilst there is evidence that combination methods often produce more accurate forecasts than individual methods (Clemen, 1989; Goodwin, 2009; Makridakis & Winkler, 1983), there is little guidance available about choosing the correct combination method and the base models to include

in the ensembles. Although forecast combination techniques such as FFORMA have recently shown competitive results in demographic forecasting (Wilson et al., 2021b), they are yet to be trialed with base models which are specifically build for small area population forecasting. Moreover, machine learning based forecasting models such as LGBM have never been evaluated on small area population datasets, despite their strong performance in the forecasting field.

The purpose of this research is to provide a comprehensive evaluation of simple and sophisticated forecast combination methods for a range of base models, including those from the demographic and time series forecasting literatures. Specifically, our paper has the following three key research aims.

1) To identify the base models that produce the most accurate small area forecasts;
2) To identify the combination methods that improve forecast accuracy and assess whether sophisticated combination methods can outperform the simpler ones;
3) To investigate how base model diversity can affect the accuracy of ensemble forecasts. We consider whether the accuracy of ensembles that only include traditional demographic models can be improved through the addition of globally trained LGBM method.

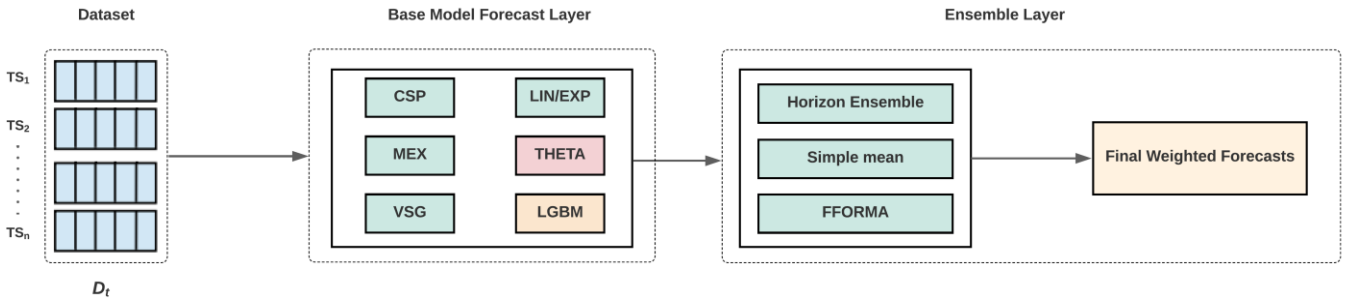
In this study, we generate population forecasts with 5-year horizons for Australian and New Zealand small areas. Whilst some applications require longer forecasts, short horizons are also widely used in demographic forecasting. A survey of Australian users of population projections found that 64% use forecasts with horizons of 5 years or less; the rest use exclusively longer forecasts (Diamond et al., 1990). We report the results for each year within the 5-year forecast window in Section 3, which is relevant to researchers and practitioners who are more interested in nowcasting. Because population estimates are often published by statistical offices some time after their reference dates, “forecasts” one or two years out from the jump-off year are often used as estimates of current populations. The implications of these results are considered in Section 4, together with suggestions of avenues for future research.

The remainder of this paper is organised as follows: Section 2 describe the base models, the combination methods, and the overall experimental design used in this study. Section 3 presents the model evaluation results, and Section 4 discuss the main findings of our work. Finally, Section 5 concludes the paper.

2. Methods and Datasets

The proposed forecasting framework to produce small area population forecasts is illustrated in Figure 1, which consists of two components, namely: 1) the base model forecast layer to generate base model predictions, 2) the ensemble layer to combine the forecasts produced from base models. In the following sections we describe the population datasets, base models, combination methods, proposed ensemble variants, and error metrics used to assess the accuracy of the population forecasts.

Figure 1: An overview of the framework to generate combination forecasts



*N*ote: The framework includes a base model forecast layer and an ensemble layer. Here, D_t denotes the target demographic dataset to forecast, where TS_1, TS_2, ..., TS_n represents the population time series relevant to n different SA2 areas. In the base model forecast layer, each model generates forecasts independently. Next, in the ensemble layer, we use different forecast combination techniques proposed in Section 2.4 to aggregate the base forecasts, and to calculate the final forecasts.

2.1. Population datasets

As the benchmark datasets, we use the mid-year Estimated Resident Population (ERP) totals for SA2 areas of Australia (Australian Bureau of Statistics, 2017) and New Zealand (Statistics New Zealand, 2020). SA2 areas comprise the smallest units of the geographical hierarchies of both countries for which population data is regularly published. Forecasts are not prepared for SA2s with population counts less than 100 in any year in the base period. We obtained 25 years' worth of yearly data available from 1991 to 2016 for 2,066 SA2 areas in Australia, and 24 yearly observations available from 1996 to 2020 for 2,053 SA2 areas in New Zealand. Australian SA2s have relatively larger populations, with a median of 9,681 in 2016, while New Zealand SA2s in 2020 had a median population of 2,340. The last 5 years of ERP time series are reserved for model testing purposes, i.e., 2012 to 2016 for Australia and 2016 to 2020 for New Zealand. Dataset characteristics are summarised in Table 1. We make these datasets publicly available through our repository for small area demographic research: https://demographic-datasets-network.github.io/.

Table 1. Characteristics of the Australian and New Zealand SA2 population forecasts

	Australia	**New Zealand**
Base Period	1991 - 2011	2001 – 2015
Forecast Period	2012 – 2016	2016 – 2020
Base Period Length (years)	21	16
Number of time series	2066	2053
Jump off year	2011	2015
Median population (jump off year)	9066	2050
Mean population (jump off year)	10735	2106
Standard deviation (jump off year)	6599	1020

2.2. Exploratory analysis

The accuracy of forecasting methods will vary depending on the characteristics of the datasets to which they are applied. Understanding these characteristics helps to provide insights into why certain methods work well for some datasets but not for others. In this section we conduct an

exploratory analysis of the similarities and differences between the Australian and New Zealand small area population time series datasets. To do this we use the R "tsfeatures" package which allows you to extract characteristics or "features" from time series data including measures of trend entropy, and stability (Kang et al., 2020; Hyndman et al., 2021a). This package has previously been used for time series analysis by data science researchers (Bojer & Meldgaard, 2021; Hewamalage et al., 2021a; Godahewa et al., 2021b).

Because we are evaluating forecasting methods, we are primarily interested in features which are most associated with the forecastability of a time series. Previous research (Bojer & Meldgaard, 2021; Hewamalage et al., 2021a; Godahewa et al., 2021b) has suggested that these features include entropy, trend, first order autocorrelation (*x_acf1*), and the optimal Box-Cox transformation parameter (lambda). The entropy feature gives an indication of the difficulty of forecasting a time series, with a higher entropy value indicating that a time series is harder to forecast. The trend is the long-term direction in which the time series moves. The first order autocorrelation feature measures the relationship between a time series and its one-step lagged series. The lambda parameter gives an indication of the variance of the time series. For more information on these extracted features, we refer to Kang et al., (2017) and Yang and Hyndman (2021).

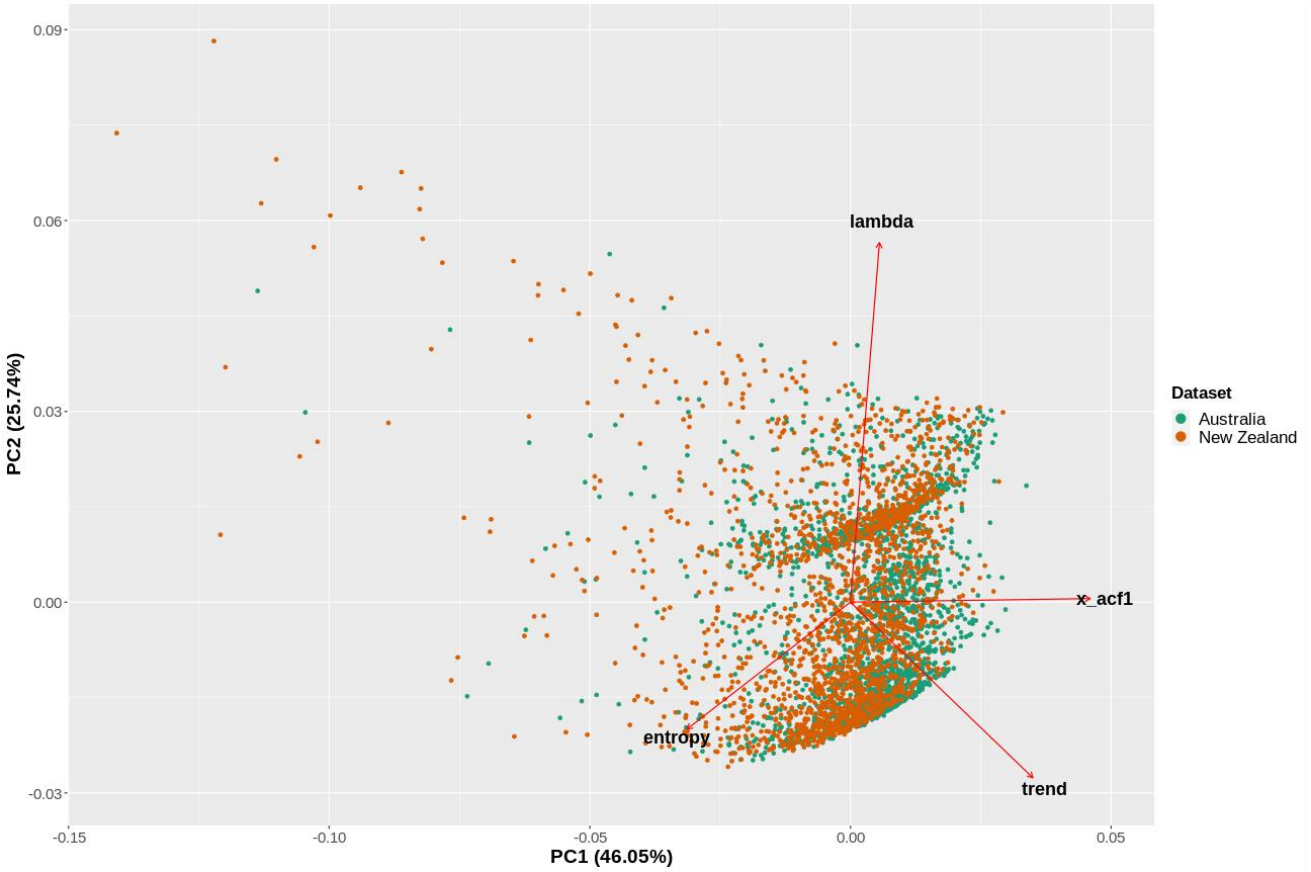
As described above, we selected four features to describe each time series. We use a dimensionality reduction algorithm, Principal Component Analysis (PCA; Jolliffe, 2011), to reduce the extracted feature space to two dimensions, so the feature distribution of the population time series can be easily visualized and compared between our two datasets. Thus, every small area time series becomes represented by two principal components, that can be plotted as a point on a graph. This plot of principal components is shown in Figure 2. Whereas the original set of four features may be correlated with each other, the principal components are independent. Each of the original features will "pull" the principal components in a particular direction, which are indicated by the red lines on Figure 2. Each point represents a small area time series, with Australian and New

Zealand time series presented in different colours.

Because both the Australian and New Zealand datasets have over 2000 points, which often overlap in Figure 2, it is difficult to compare the two datasets. We address this by plotting the density of the points of the points generated through PCA using Hexbin plots. Figure 3 shows hexbin plots of the normalised density values for Australian and New Zealand population datasets.

Figure 3 shows that the feature spaces of the two population datasets are highly populated on the bottom-right side and middle-right side. This suggests that both datasets contain many time series with high levels of trend and first order correlation, meaning many small areas in both Australia and New Zealand are growing in population size and population counts are strongly correlated with those in the preceding year. It is noticeable that the New Zealand dataset contain more high-density areas compared to Australian dataset. This indicates that the New Zealand dataset has a higher number of similar time series.

Figure 2: Tsfeature plots for the Australian and New Zealand Datasets



*N*ote: The directions of the 4 feature components: spectral entropy (entropy), strength of trend (trend),first order autocorrelation (*x_acf*1) and optimal Box-Cox transformation parameter (lambda) generated bythe first two principal components (PC1, PC2) are represented by red directed arrows. The time series in the two datasets are differentiated by the colour of the plotted dot.

Figure 3: Hexbin plots for the Australian and New Zealand Datasets

*N*ote: The dark and light bin hexbins represent the high and low feature density areas.

2.3. Base Models

In our chosen framework, we consider both univariate and global forecasting models to incorporate the strengths of different model classes. In small area population forecasting, this strategy can be useful for modelling different types of population growth patterns. For example, univariate models can capture the population dynamics of a single area, whereas global models can be used to capture population growth patterns common across multiple areas. According to Lichtendahl and Winkler (2020) model diversity can be useful in producing better results from a forecast ensemble.

The proposed forecasting ensemble contains six base models, including four univariate forecast models which have previously been shown to produce accurate estimations for small area populations. These models are the Constant Share of Population (CSP; Smith et al., 2016), Linear/Exponential (LIN/EXP; Wilson, 2015), Modified Exponential (MEX; Baker et al. 2008), and Variable Share of Growth (VSG; Wilson, 2015) models. Additionally, we include THETA, a univariate forecast model suitable for yearly time series forecasting, and a globally trained LGBM model that exploits the similarities between a collection of population time series. These base models are briefly summarised in Table 2. More specifically, the demographic models are described briefly in Section 2.3.1, the THETA model is described in Section 2.3.2, and the LGBM model is described in Section 2.3.3.

Table 2: The base models used in our chosen framework

Model	Summary
CSP (Smith et al., 2013)	• Assumes a small area's share of the national or state population stays constant. • Use national/state forecast and each area's share of the national population in the jump-off year to create small area forecasts.
LIN/EXP (Wilson, 2015)	• Linear model if base period population growth is positive; exponential method if growth is negative. • Prevents runaway growth for rapidly growing areas, and negative forecasts for rapidly declining populations.
VSG (Wilson, 2015)	• LIN/EXP forecast is adjusted so that forecast growth summed over all small areas equals national growth.
MEX (Baker et al., 2008)	• Exponential forecast, modified with a floor or ceiling limit to prevent runaway growth or improbably low forecasts.
THETA (Assimakopoulos & Nikolopoulos, 2000)	• A special case of the simple exponential smoothing with drift method.
LGBM (Ke et al., 2017)	• A globally trained LGBM machine learning model. • Time series data from all SA2 areas are used to build a unified forecasting model.

2.3.1. Demographic Models

In this section, we briefly describe the four demographic models used in this study. For further discussions and definitions of these methods, we refer to Wilson (2015). In general, these traditional demographic models are simple, easy to implement, readily interpretable, and require little data. The LIN/EXP and MEX methods produce forecasts using only the population totals for the jump-off year and those from 10 years prior; the CSP and VSG methods disaggregate national (or State) population forecast to small areas (Australian Bureau of Statistics, 2008, 2013; Statistics New Zealand, 2009, 2014). Here, jump-off year refers to the last year of the base or fitting period.

The LIN/EXP method projects future populations using a linear model if the population growth over the base period is positive, whereas if the growth rate over the base period is negative, it uses an exponential method. In this way, it prevents the linear model from creating "negative" populations for small areas with rapidly declining populations and the runaway growth produced by the exponential method for areas with rapid population growth. This model was one of the top performing methods identified in Wilson (2015), and a similar composite model evaluated by Rayer (2008) was shown to produce accurate small area forecasts.

Runaway growth produced by the exponential method for areas with very high growth rates over the base period can be dampened with a modification used by Baker et al. (2008). The equation for the MEX model is dependent on whether the population is increasing or decreasing during the base period.

If the population is increasing during the base period

$$P_i(t+1) = P_i(t)e^{[G_i\left(1-\frac{P_i(t)}{K_i}\right)]}$$

If population is decreasing during the base period

$$P_i(t+1) = P_i(t)e^{[G_i\left(1-\frac{K_i}{P_i(t)}\right)]}$$

where $P_i(t)$ is the population of small area i in the jump-off year, G_i is the average annual growth over the base period and the damping factor K_i. In the population growth scenario, K_i is small area i's maximum population, whereas in the population decreasing scenario it is the minimum population. Ideally K_i would be calculated based on a range of area specific characteristics, such as historic land use zoning regulations. However, such data is often unavailable. The key purpose of K_i is to prevent runaway growth. This is achieved here by setting K_i to 5, following Wilson (2015). Although to some extent arbitrary, this prevents the forecast population from increasing beyond five times the jump off population or decreasing below one fifth its value. The need to set K_i using external information, be it external variables or practitioner judgement, is a limitation of the MEX method.

The CSP method assumes that a small area's share of the national (or State) jump-off population remains constant over time. The forecast is created by calculating each small area's share of the national population in the final year of the base period (the last year of the training data), and then projecting future populations as a constant share of the national forecast.

The Variable Share of Growth model allocates forecast national (or State) population growth to small areas (Smith et al. 2013). In our application, we first create preliminary forecasts of growth for small areas using the LIN/EXP model. These are then modified using the plus-minus method (Shyrock & Siegel, 1973) so that forecast growth across all small areas equals national growth from the independent forecast.

2.3.2. THETA

THETA is a univariate, non-seasonal forecast model that extrapolates the trend and the level of a series (Assimakopoulos & Nikolopoulos, 2000). In the forecasting literature, the THETA model is also treated as a special case of the simple exponential smoothing drift (SES) method (Hyndman & Billah, 2003). It was the best performing model in the M3 competition (Makridakis & Hibon, 2000), and is also included as a base model in the original FFORMA ensemble framework that achieved the second best results in the M4 competition (Makridakis et al., 2018). Therefore, we include THETA as a good representative model to forecast small area populations datasets that mostly contain non-seasonal, yearly time series. In our experiments, the THETA method is used from the "thetaf" function from the R package "forecast" (Hyndman & Khandakar, 2008; Hyndman et al., 2021b).

2.3.3. Light Gradient Boosting Model (LGBM)

LGBM models (Ke et al., 2017) are a popular and computationally efficient machine learning algorithm which are a variant of Gradient Boosting Models (GBM). GBM-based algorithms use training data from the base period to learn the rules for a decision tree which assigns an output value to a given time series input. For example, if the data in the base period includes the sequences [1 2 3] and [3 4 5], the tree can create a set of rules that assigns an output of 3 to an input sequence of [1 2]. Of course, real world data is much more complicated and the first decision tree that is built is usually quite poor at correctly assigning an output value. GBMs use the performance of the first decision tree to create an improved decision tree. This process is repeated sequentially, and the outputs are accumulated to form the final output of the model. Therefore, the initial "weak" decision tree is "boosted" to produce more accurate predictions. We use LGBM as our main non-linear model in our proposed ensemble framework. LGBM is a highly efficient GBM variant that has shown promising performance in the recent M5

forecasting competition (Makridakis & Spiliotis, 2021). In our experiments, we use the implementation available in the lightbm function from the Python package "lightbm" (Ke et al., 2021).

When implementing global models it is important to account for the various scales and variance present in the pool of time series as they are trained across a group of time series. For example, in the demographic forecasting context, the scales of the population levels in small areas can be different (e.g., small area populations can range between the 100s and 10, 000s). Therefore, as a preprocessing step, we first normalise the set of target time series ($\chi = \{X_i\}_{i=1}^{n}$) using a meanscale transformation strategy (Bandara et al., 2020a; Hewamalage et al., 2021b), which can be defined as follows:

$$x_{i,normalised} = \frac{X_i}{\frac{1}{k}\sum_{t=1}^{k} X_{i,t}}$$

where $X_{i,\text{normalised}}$ represents the ith normalised time series, and k is the number of observations in time series X_i, where $i \in \{1, 2, \cdots, n\}$. To train the LGBM model, we use the normalized observations of time series in the form of input and output frames. We achieve this by applying the Moving Window (MW) transformation strategy to each time series. Then, following the recommendations of (Hewamalage et al., 2021b), a global model is built by pooling these input and output frames together. The default implementation of LGBM models is only able to train a single output from a given set of inputs. On the other hand, neural network models are multi-output regression models, which can train multiple outputs at the same time. Because LGBM models are single-output regression models, they can only generate one-step-ahead predictions. As a result, we set the size of the output window to one. Similar to Bandara et al. (2021b), a recursive forecasting strategy is used to generate forecasts for multiple steps

ahead. This involves feeding the prediction from the last time step as input for the next prediction. In addition to the past observations, we use the mean and the standard deviation of each input frame as exogenous variables to train the LGBM.

Table 3: The hyper-parameter ranges used to train LGBM in our experiments

Model Parameter	Minimum value	Maximum value
Learning rate	0.025	0.090
Sub row	0.70	0.90
L2-regularisation weight	0.1	0.3
Number of iterations	1000	2000
Number of trees	200	400

Our LGBM contains various hyper-parameters, including learning rate, sub row, model regularization terms, , number of iterations, and number of trees. We use Hyperopt (Bergstra et al., 2013), a python implementation of a Bayesian hyper-parameter optimization process, to autonomously determine the optimal values for these hyper-parameters of LGBM. Table 3 summarises the ranges of hyper-parameter values explored in our experiments.

When we train machine learning models there is a risk that our models become too specific for the training data, and not generalizable to unseen data. This phenomenon is referred to as "model overfitting" in the machine learning literature. To minimise model overfitting, practitioners set aside a portion of the training data, known as "validation data", which is used to evaluate how model performs on the unseen data. There are multiple ways to split data from the base period into training and validation sets.

Figure 4: The training and validation split of time series

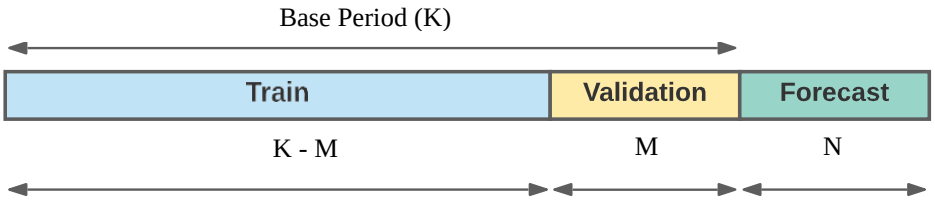


Figure 4 shows the splitting criteria used in our experiments. Given a length of time series K, we reserve M number of observations from the end of the time series for model validation in the hyper-parameter optimisation procedure. Here, $K - M$ observations are used to train LGBM, where the size of M is equivalent to the length of the intended forecast horizon N, i.e., $M = N = 5$ in our experiments. This is in line with the recommendations given by Hewamalage et al. (2021b) and Bandara et al. (2020b) for global model training. To address the parameter uncertainty of LGBM during model training, we train the LGBM on 5 different seeds using the optimised hyperparameters from Hyperopt and compute the median of the predictions produced from different initialization seeds.

2.4. Combination Methods

In this section we describe the details of the four forecast combination methods used in this study. The combination methods used in our study are summarized in Table 4 and detailed descriptions are provided in the sections below.

2.4.1. Average-based combination methods

A simple mean of forecasts is considered as one of the most computationally efficient combination forecasting methods that often produces competitive results (Genre et al., 2013; Shaub, 2020; Timmermann, 2006). Additionally, trimmed means have been suggested in the demographic literature to avoid extremely high or low forecasts affecting the mean value (Rayer et al., 2009; Rayer & Smith, 2010). Therefore, we evaluate both simple mean (MEAN) and trimmed mean (TRIMMED) combination benchmarks.

2.4.2. Feature-based model averaging

We use a feature-based forecast combination framework, FFORMA, introduced by Montero-Manso et al. (2020), to combine the forecasts produced from our base models. The original implementation of FFORMA uses nine statistical forecasting methods from the "forecast" package (Hyndman & Khandakar, 2008; Hyndman et al., 2021b). In simple terms, FFORMA produces a weighted combination forecast. The weights for each of the included base models are obtained by a learning model based on a gradient boosted tree trained using 42 time series specific features, including trend, seasonality, entropy, and linearity. These time series features are calculated using the R package "tsfeatures" (Hyndman et al., 2021a).

In the training phase, FFORMA first splits each time series into training validation sets. A validation set is taken from the end of each time series and its size is equivalent to the intended forecast horizon. That is, the last 5 years of the data in the base period is separated as a "validation set" from the earlier time points in the time series. Secondly, time series features are calculated using the training set. This involves generating forecasts for the last 5 years of the base period. Thirdly, FFORMA calculates the forecast errors for each base model on the validation set. This involves evaluating the error of the forecasts for the last 5 years of the base period. Next, a gradient boosted tree is trained using the extracted time series features and the base model errors. Here, the time series features are used as the input values of the model, whereas the base model errors on the validation set are considered as the output values of the model. This trained gradient boosted tree assigns a set of weights for each base model that can be used to combine the individual base forecasts generated in the final stage. In the testing phase, FFORMA calculates the time series features using the entire time series and base model forecasts are generated for the expected forecast horizon. As the last step, FFORMA combines these forecasts using the optimal set of weights obtained by the gradient boosted tree in the training phase. For more detailed discussions of the FFORMA meta-learning methodology, see

Montero-Manso et al. (2020).

In our work we replace the default base models used in the original implementation of FFORMA with CSP, LIN/EXP, MEX, VSG, THETA, and LGBM. We use the FFORMA implementation available from the M4 metalearning package (Montero-Manso et al., 2020) in our experiments.

2.4.3. Horizon based combination

We introduce an ensemble that combines forecasts of the best performing base models of each forecast horizon. This consists of creating forecasts for the last 5 years of the base period and, for each year, selecting the three top performing methods. These methods are then used to produce the required forecast for the intended forecast horizon. So, if the CSP, VSG and LGBM models are the most accurate models for forecasting the last year in the base period, then a combination of these methods is used to create the forecast for 5 years after the base period. More specifically, the horizon-based combination involves the following steps. Following the train-validation split procedure used in FFORMA, we first create a training and a validation set for each time series. Next, the base model forecasts are generated for the validation time period using training data. On the validation set, we calculate the mean forecast error of the base models for each horizon, and select horizon-wise the top three performing models with the smallest errors. In the testing phase, after obtaining the base model forecasts using the full time series, we calculate the horizon-wise mean of the forecasts provided by the best-performing models identified in the training phase. In this way, we aim to develop a specialised forecast ensemble for each forecast horizon.

2.5. Benchmarks and Model variants

In addition to the six base models introduced in Section 2.3, we use ES-RNN (Smyl, 2019) and FFORMA (Montero-Manso et al., 2020), the two best-performing methods in the M4 forecasting competition, as benchmarks to compare against our proposed ensemble variants. In our experiments, we refer the results of the original implementation of FFORMA as FFORMA-ORIGINAL. The benchmarks and base models are compared against the ensemble model variants in Table 4. These variants are based on applying the described combination methods to the chosen ensembles.

Table 4. Considered ensembles, combination methods, and Ensemble Model Variants

Ensembles	Included Base Models
ALL	CSP, LIN/EXP, MEX, VSG, THETA, and LGBM
POPEXPERT	CSP, LIN/EXP, MEX, and VSG
POPEXPERT-GLOBAL	CSP, LIN/EXP, MEX, VSG, LGBM
Combination Method	**Description**
MEAN	Simple mean of forecasts from all models in the combination forecast
TRIMMED	Mean of constituent forecasts after highest and lowest predictions removed
FFORMA	Considers the performance of models over the base period to weight forecasts.
HORIZON	Identifies the top 3 performing methods for forecasts in the base period and use their combinations to forecastfuture values
Ensemble Model Variants	**Description**
MEAN-ALL	The simple mean of the forecasts of the six base models: CSP, LIN/EXP, MEX, VSG, THETA, and LGBM
MEAN-POPEXPERT	The simple mean of the forecasts from the four demographic base models: CSP, LIN/EXP, MEX, VSG
MEAN-POPEXPERT-GLOBAL	The simple mean of the forecasts from the four demographic base models and the global forecasting model: CSP, LIN/EXP, MEX, VSG, LGBM
TRIMMED-ALL	The trimmed mean of the forecasts from all six base models: CSP, LIN/EXP,MEX, VSG, THETA, and LGBM
TRIMMED-POPEXPERT	The trimmed mean of the forecasts provided by the four demographic base models: CSP, LINEEXP, MEX, VSG
FFORMA-ALL	The modified version of FFORMA which uses the six models as its base models: CSP, LIN/EXP, MEX, VSG, THETA, and LGBM
FFORMA-POPEXPERT	The modified version of FFORMA which uses the four demographicbase models: CSP, LIN/EXP, MEX, VSG
FFORMA-POPEXPERT-GLOBAL	The modified version of FFORMA which uses the four demographic base models together with the global forecasting model, as its base models: CSP, LIN/EXP, MEX, VSG, and LGBM.

Note: The POPEXPERT-GLOBAL model is included to evaluate the benefit of adding LGBM to a pool of specialised demographic models. We cannot know in advance which base models will be included in a trimmed mean. Therefore, we do not include a TRIMMED-POPEXPERT-GLOBAL ensemble model variant as it is not suitable to evaluate the impact of including LGBM.

2.6. Error Metrics

We assess the performance of the proposed framework using absolute percentage error (APE), which is a relative error measure commonly used in the population forecast literature (Rayer, 2007). The APE at time *t* is defined as follows:

$$APE = \frac{|F_t - A_t|}{A_t} * 100$$

Here, A_t represents the actual population at time *t*, and F_t is the population forecast generated by a prediction model. To summarise the overall APE for each prediction horizon, we use the Median Absolute Percentage Error (MedAPE). The APE distributions generated from population forecasts are typically negatively skewed with a long tail of errors, therefore we choose to report median APE over mean APE to reduce the impact of extreme errors.

MedAPE allows us to evaluate performance across time series, however it is possible that some methods will show good MedAPEs, but have many relatively poor forecasts, and vice versa. To indicate the distribution of errors we introduce the Absolute Percentage Error Rank (APE Rank) metric. To calculate this metric, each of the forecasts is ranked by their absolute percentage error for each of the small areas. As a total of 17 forecasts are being evaluated (including those produced by individual models, ensemble models, ES-RNN and FFORMA-ORIGINAL), the rankings vary from 1 to 17, where a rank of 1 indicates the top performing method. This will generate a set of rankings for each small area by each method. Therefore, each method will have 2,066 APE ranks for Australian SA2s and 2,053 APE ranks for New Zealand SA2s. The APE Rank distributions are visualized using violin plots in our analysis.

2.7. Statistical test of the results

To evaluate the statistically significant differences within the proposed ensemble variants and base models, we use the non-parametric Friedman rank-sum test. We use Hochberg's

post-hoc procedure to further examine these differences (Garcia et al., 2010). The APE error measure is used to perform the statistical testing, with a significance level of $\alpha = 0.05$.

3. Results

This section details the overall results and compares the performance of the base models and the proposed ensemble variants for the Australian and New Zealand small area population datasets. Here we present the results for forecast horizons of 1, 2, 3, 4, and 5 in tabular form, and violin plots and statistical tests for a forecast horizon of 5. The rest of the violin plots and statistical tests, in terms of the APE rank for forecast horizons of 1, 2, 3 and 4 years are available in Appendix A and B, respectively.

3.1 LGBM Global vs LGBM Univariate

First, we highlight the suitability of training a LGBM model globally across many yearly population time series, instead of training multiple LGBM models for each time series separately. Apart from the training methodology, i.e., global vs univariate, the experimental settings of LGBM-Global and LGBM-Univariate are identical as described in Section 2.3.3

According to Table 5 and Table 6, it can be seen that the proposed LGBM-Global variant outperforms the LGBM-Univariate model consistently across all the horizons. The substandard performance of the LGBM-Univariate model can be attributed to the limited availability of data in a single population time series. As the LGBM-Univariate model trains multiple LGBM models for each time series separately, LGBM models are unable to fit their model parameters accurately with such a short time series. On the other hand, LGBM-Global model is trained across the entire

population dataset, hence has access to many population time series to estimate the model parameters accurately.

Therefore, in our experiments, we use the LGBM-Global variant as our primary machine learning base model in our proposed ensembles.

Table 5: MedAPE of LGBM-Global and LGBM-Univariate models for the Australian small area forecasts

	Forecast Horizon (years)				
Method	**1**	**2**	**3**	**4**	**5**
LGBM-Global	0.702	1.382	2.100	2.910	3.842
LGBM-Univariate	7.640	8.740	9.490	10.325	11.072

Table 6: MedAPE of LGBM-Global and LGBM-Univariate models for the New Zealand small area forecasts

	Forecast Horizon (years)				
Method	**1**	**2**	**3**	**4**	**5**
LGBM-Global	0.698	1.261	1.841	2.369	3.092
LGBM-Univariate	7.405	8.649	9.903	10.917	12.313

3.2 Australian small area forecasts

Table 7 summarises the overall results of the proposed variants and the benchmarks for each forecast horizon, in terms of the MedAPE metric. According to Table 7, the proposed MEAN-ALL variant achieves the best MedAPE for forecast horizons of 1, 2, 3, and 4 years, whereas the proposed TRIMMED-ALL variant obtains the best MedAPE for a forecast horizon of 5 years. With respect to individual models, we see that LIN/EXP model performs the best, followed by MEX and VSG. Among average-based ensembles, we see that MEAN-POPEXPERT-GLOBAL gives better results compared to MEAN-POPEXPERT. This indicates that the performance of MEAN-POPEXPERT improves after adding the global forecasting base model, LGBM, to the

model combination. This observation can also be seen among the feature-based ensemble models, where FFORMA-POPEXPERT-GLOBAL obtains better MedAPEs across all prediction horizons, compared to FFORMA-POPEXPERT, which only contains the demographic base models. We also see that the proposed FFORMA variants are unable to outperform the FFORMA-ORIGINAL benchmark.

Table 7: MedAPE of the Australian small area population forecasts

	Forecast Horizon (years)				
Method	**1**	**2**	**3**	**4**	**5**
CSP	1.193	2.390	3.642	5.015	6.183
LIN/EXP	0.672	1.307	1.899	2.583	3.302
MEX	0.678	1.312	1.948	2.631	3.319
VSG	0.678	1.317	2.032	2.770	3.401
THETA	0.827	1.614	2.391	3.048	3.852
LGBM	0.702	1.382	2.100	2.910	3.842
MEAN-ALL	**0.600**	**1.175**	**1.787**	**2.435**	3.129
MEAN-POPEXPERT	0.636	1.262	1.935	2.693	3.422
MEAN-POPEXPERT-GLOBAL	0.618	1.208	1.853	2.507	3.217
TRIMMED-ALL	0.620	1.202	1.831	2.441	**3.094**
TRIMMED-POPEXPERT	0.664	1.291	1.909	2.571	3.294
FFORMA-ALL	0.691	1.385	2.086	2.840	3.592
FFORMA-POPEXPERT	0.804	1.581	2.417	3.328	4.153
FFORMA-POPEXPERT-GLOBAL	0.684	1.378	2.098	2.855	3.609
HORIZON-ALL	0.617	1.184	1.821	2.449	3.278
ESRNN	0.653	1.789	2.981	4.232	4.863
FFORMA-ORIGINAL	0.684	1.335	1.998	2.651	3.430

*N*ote: Results are reported for all years within the 5 year forecast horizon for the 2066 Australian SA2 areas. For each forecast horizon, the results of the best performing method(s) are marked in boldface. The base forecast models, average-based ensemble models, the feature-based ensemble models, the horizon-based ensemble models, and the additional benchmarks are separated with horizontal lines.

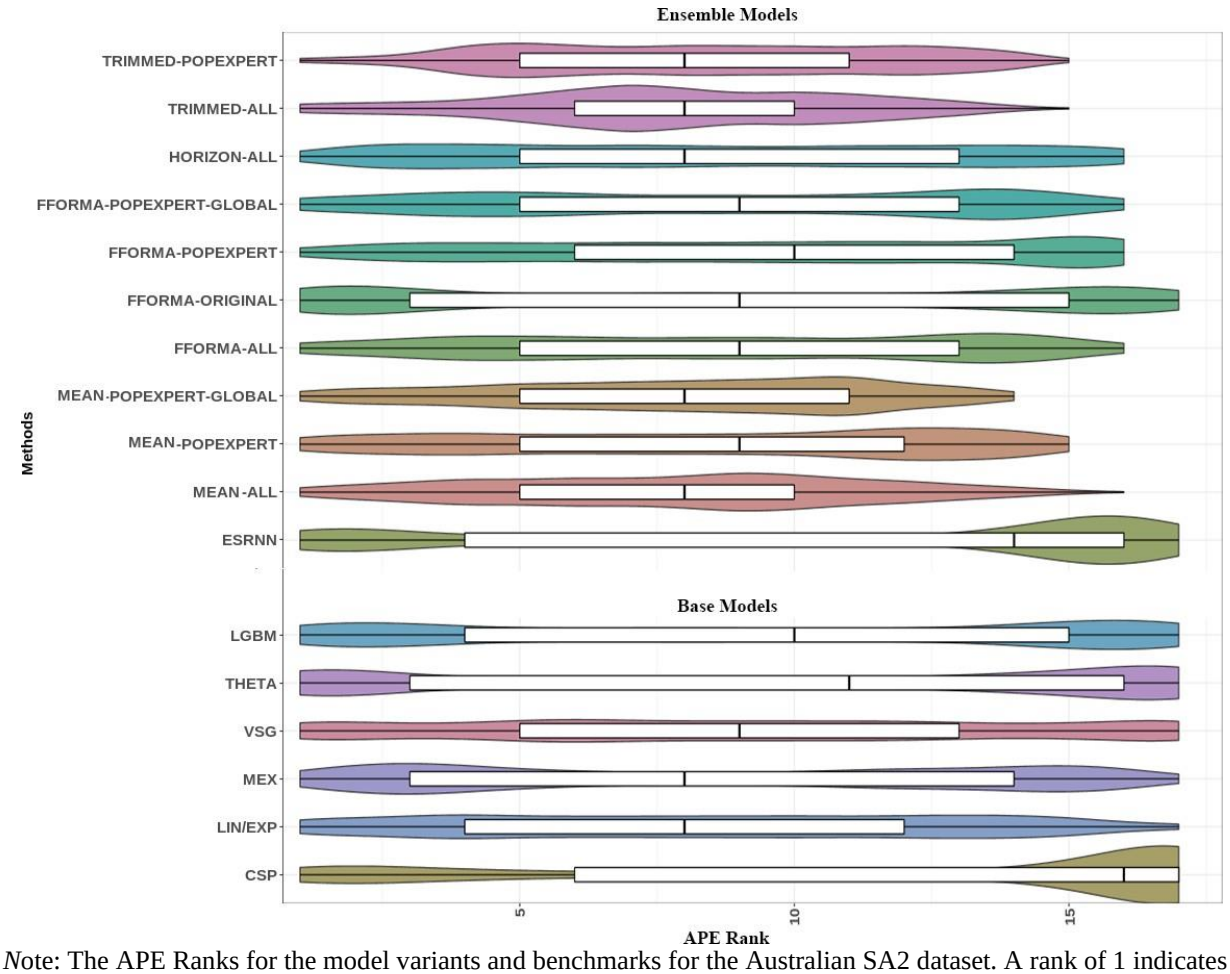
These visualisations allow us to address our three research aims. First, we show that the proposed combination methods generally – but not always – performed better than the individual forecasts, however the worst forecast by an individual model has a markedly greater MedAPE than the worst ensemble forecast. Then, we show that ensembles containing LGBM as a base

model generally outperform those without LGBM model. Finally, we can observe that in most cases, the average-based combination models, which include MEAN and TRIMMED methods, are able to outperform more complex feature-based and the horizon-based ensemble models.

Finally, the distribution of APE ranks for year 2016 are shown in Figure 5 using violin charts. Here, the models are ranked by their absolute percentage error for each small area. Each violin plot visualizes the distribution of APE ranks for the method indicated on the vertical axis. The horizontal axis displays the APE Rank, where the violin plot is wide at the beginning of the axis, this indicates that the models are top performers for many of the small areas. Where the violin plot is wide at the tail, this indicates that the model is amongst the worst performers for many of the areas. Each violin plot contains a box plot, where the edges of the boxes represent the 25th and 75th quartiles, and the vertical bar represents the median.

According to Figure 5, we see that MEAN-ALL, MEAN-POPEXPERT-GLOBAL, and HORIZON-ALL model variants shows the best APE ranks among the ensemble model variants. Moreover, the proposed ensemble models tend to have fewer relatively bad forecasts when compared to individual methods, even if they may not have more top ranked forecasts. This is indicated by the smooth tails on the violin plots for the ensemble models. The individual base models tend to have more poorly ranked forecasts, particularly the CSP and ES-RNN methods. The best performing base models are MEX and LIN/EXP. Figure 5 shows that whilst the MEX and LIN/EXP models produce forecasts with a similar MedAPE ranks, the LIN/EXP model has a fewer top and bottom APE ranks than the MEX model. Whereas, the VSG model has a relatively flat distribution, indicating that it produces forecasts that are both good and bad.

Figure 5: The APE Ranks for the model variants and benchmarks for the Australian SA2 dataset



*N*ote: The APE Ranks for the model variants and benchmarks for the Australian SA2 dataset. A rank of 1 indicates the top performing method, whilst a rank of 17 indicates the worst performing method. There are 2,066 rankings for each model, one for each Australian small area.

Table 8 shows the results of statistical testing for the APE error measure for the 5th forecast horizon of the Australian dataset. Adjusted p-values calculated from the Friedman test with Hochberg's post-hoc procedure are shown. The overall result of the Friedman rank sum test for Australian dataset is a p-value of $p < 10^{-10}$, which means the results are highly significant. The table shows that the proposed TRIMMED-ALL variant performs the best and is used as the control method. Moreover, the MEAN-ALL and MEAN-POPEXPERT-GLOBAL

variants do not perform significantly worse than the control method. Most importantly, it can be observed that the TRIMMED-ALL method achieves significantly better results than the individual models.

Table 8: The Friedman rank sum test for the Australian small area population dataset.

Method	P_{Hoch}
TRIMMED-ALL	-
MEAN-ALL	0.859
MEAN-POPEXPERT-GLOBAL	0.269
TRIMMED-POPEXPERT	1.740×10^{-3}
LIN/EXP	1.865×10^{-5}
MEX	3.443×10^{-6}
MEAN-POPEXPERT	2.704×10^{-6}
HORIZON-ALL	1.151×10^{-6}
FFORMA-POPEXPERT-GLOBAL	6.419×10^{-9}
FFORMA-ALL	2.400×10^{-9}
FFORMA-ORIGINAL	3.088×10^{-12}
VSG	2.912×10^{-16}
LGBM	1.687×10^{-25}
FFROMA-POPEXPERT	3.126×10^{-31}
THETA	1.866×10^{-38}
ESRNN	4.317×10^{-76}
CSP	2.011×10^{-168}

Note: A horizontal line is used to separate the methods that perform significantly worse than the best performing method.

3.3 New-Zealand small area forecasts

The results for the evaluation of forecasting methods for New-Zealand small area populations are shown in Table 9. It shows that the global forecasting base model LGBM achieves the best performance both amongst the individual models and overall. It also outperforms the ES-RNN and FFORMA-ORIGINAL benchmarks. Despite the standout performance of LGBM, ensemble models generally outperform individual models. The more complex HORIZON-BASED and FFORMA-BASED combination methods outperform the ensemble forecasts generated from the MEAN and TRIMMED combination methods.

Furthermore, the proposed feature-based ensemble models FFORMA-ALL, FFORMA-POPEXPERT, and FFORMA-POPEXPERT-GLOBAL obtain better MedAPE results compared to the FFORMA-ORIGINAL variant.

Table 9: MedAPE of the New Zealand small area population forecasts

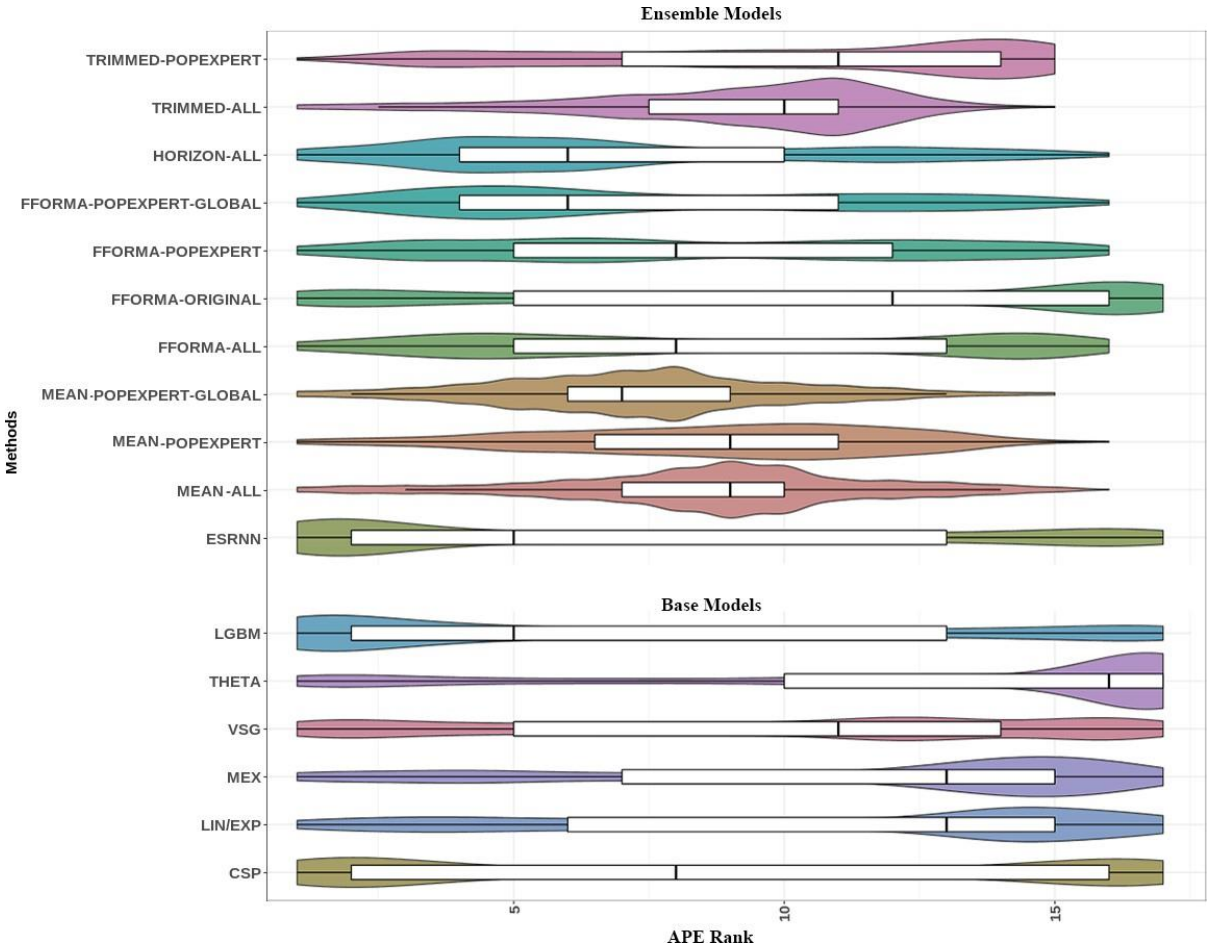
	Forecast Horizon (years)				
Method	**1**	**2**	**3**	**4**	**5**
CSP	0.862	1.624	2.412	3.203	3.767
LINEEXP	1.101	2.041	2.784	3.521	4.356
MEX	1.154	2.103	2.894	3.618	4.539
VSG	1.065	1.935	2.721	3.436	4.317
THETA	1.396	2.520	3.492	4.420	5.498
LGBM	**0.698**	**1.261**	**1.841**	**2.369**	**3.092**
MEAN-ALL	0.981	1.728	2.445	3.053	3.871
MEAN-POPEXPERT	0.977	1.764	2.505	3.125	3.937
MEAN-POPEXPERT-GLOBAL	0.909	1.641	2.272	2.928	3.696
TRIMMED-ALL	1.015	1.813	2.479	3.163	4.004
TRIMMED-POPEXPERT	1.081	1.994	2.736	3.463	4.359
FFORMA-ALL	0.960	1.677	2.356	3.109	3.818
FFORMA-POPEXPERT	0.912	1.649	2.364	3.086	3.653
FFORMA-POPEXPERT-GLOBAL	0.846	1.543	2.216	2.933	3.520
HORIZON-ALL	0.963	1.865	2.606	2.883	3.274
FFORMA-ORIGINAL	1.095	2.000	2.764	3.513	4.434
ESRNN	0.994	1.729	2.127	2.729	3.259

*N*ote: Results are reported for all years within the 5 year forecast for the 2053 New Zealand SA2 areas. For each forecast horizon, the results of the best performing method(s) are marked in boldface. The base forecast models, average-based ensemble models, the feature-based ensemble models, the horizon-based ensemble models, and the additional benchmarks are separated with horizontal lines

Similar to Figure 5, the distribution of APE ranks for each of the models for New Zealand dataset, are shown in Figure 6 using violin plots. The violin plots correspond to a 5 year forecast horizon. According to Figure 6, it can be seen that the LGBM and ESRNN methods are the top performers, with a high density of top ranked forecasts, and relatively few bottom ranked forecasts. The VSG model has a relatively flat distribution. The MEX and LIN/EXP methods perform poorly with a high density of bad forecasts. The MEAN and HORIZON-based

combination methods have notably finer tails than the FFORMA-based methods, indicating that they give fewer relatively bad forecasts, even though they do not produce lower overall error. The POPEXPERT-GLOBAL ensembles are left skewed compared to POPEXPERT ensembles, indicating that the inclusion of LGBM increases the overall number of good forecasts.

Figure 6: The APE Ranks for the model variants and benchmarks for the New Zealand SA2 dataset



*N*ote: The APE Ranks for the model variants and benchmarks for the New Zealand SA2 dataset. A rank of 1 indicates the top performing method, whilst a rank of 17 indicates the worst performing method. There are 2053 rankings for each model, one for each New Zealand small area.

Table 10: The Friedman rank sum test for the New Zealand small area population dataset.

Method	P_{Hoch}
HORIZON-ALL	-
LGBM	0.919
ESRNN	0.303
MEAN-POPEXPERT-GLOBAL	3.122×10^{-2}
FFORMA-POPEXPERT-GLOBAL	3.232×10^{-3}
FFROMA-POPEXPERT	6.679×10^{-20}
MEAN-ALL	1.273×10^{-26}
FFORMA-ALL	4.659×10^{-28}
MEAN-POPEXPERT	8.783×10^{-29}
CSP	3.605×10^{-29}
TRIMMED-ALL	2.789×10^{-40}
VSG	8.675×10^{-61}
TRIMMED-POPEXPERT	1.035×10^{-86}
FFORMA-ORIGINAL	1.140×10^{-102}
LIN/EXP	2.680×10^{-117}
MEX	8.589×10^{-157}
THETA	6.930×10^{-285}

Note: A horizontal line is used to separate the methods that perform significantly worse than the best performing method.

We observe that the overall p-value of the Friedman rank sum test for the APE error measure for the 5th forecast horizon of the New Zealand dataset is a p-value of $p < 10^{-10}$, which means the results are highly significant. According to Table 10, the HORIZON-ALL variant performs the best and is used as the control method. It can be seen that the LGBM and ESRNN methods do not perform significantly worse than the HORIZON-ALL variant.

4. Discussion

In this study, we sought to contribute to the literature by investigating how state of the art forecasting and model combination techniques can be integrated with the proven demographic models to improve forecast accuracy in the small area population forecasting field.

Our first aim was to evaluate the performance of base models for small area population forecasting. The results indicate that the proposed machine learning based forecasting technique LGBM obtains the best results for the New Zealand small area dataset. However, it was shown that the standard demographic base models are able to outperform the LGBM model on the Australian small area dataset. This variability of model performance across different datasets is a common issue pertinent to both machine learning and statistical models. LIN/EXP, MEX, VSG and THETA produced more accurate forecasts for Australia, however LGBM and CSP performed better for the New Zealand dataset. In such situations, it is important to understand the causes of such variability. First, we note that the Australian SA2s have a median jump off year population of 9066, versus 2050 for New Zealand. It is well established that forecast error increases as population decreases (Wilson et al., 2018). Thus, it is expected that forecast accuracy would be more accurate for Australia. It is more interesting that the CSP and LGBM methods do better for New Zealand. The CSP method projects small area populations such that they are forecast to grow in line with projected national growth rates from an external forecast. The better performance of CSP for New Zealand suggests there was less variability in small area population growth rates than in Australia. The performance difference of globally trained LGBM model can be explained by the discussions highlighted by Bandara et al. (2021, 2020b) and Hewamalage et al. (2021b). These studies show that the competitiveness of global models can be affected by the homogeneous characteristics present in a time series dataset. If the set of time series exhibit highly heterogeneous characteristics, training a single model across these time series can

degenerate the overall accuracy of the model. Thus, based on the performance of LGBM, it is evident that population change across small areas of New Zealand is more homogeneous compared to that for Australia. This is supported by our exploratory analysis of the Australian and New Zealand population datasets in Section 2.2 where we found that the New Zealand dataset has a higher number of similar time series. Several alternatives have been proposed for GFMs to produce more accurate forecasts when a dataset is more heterogeneous, such as time series clustering (Bandara et al., 2021) and developing ensembles of localised models (Godahewa et al., 2021). Further research is required to establish, whether use of these methods can help to achieve consistently good forecasts.

The second aim was to identify the best performing combination method for population forecasting. We observe that the simple, average-based combination methods (MEAN and TRIMMED) are able to produce more accurate forecasts for Australian small areas, whilst the feature-based combination method performs better for the New-Zealand dataset, i.e., FFORMA. Meanwhile, the horizon-based combination method, i.e., HORIZON, performs comparably well across both datasets. According to APE Rank metric, we see that the simple mean method generally produces forecasts with less variance and fewer poor forecasts than the FFORMA variants; this is true for both datasets. Furthermore, the proposed ensemble models show less variance and fewer poor forecasts than the individual models, even though they do not always outperform individual models. As advised by Hibon and Evgeniou (2005), the benefit of combining forecasts is to decrease the risk of producing a highly erroneous forecast rather than the best possible forecast. For practicing demographers, avoiding bad forecasts is often more important than modestly reducing overall error.

As the third research aim, we investigated how model diversity can affect the overall ensemble model accuracy. For both Australian and New Zealand datasets we observe the

performance gain achieved by all ensemble variants after adding LGBM as a base model. This highlights the benefit of including a global model as a base model to exploit the similar population growth patterns in a dataset. Next, we identify that the FFORMA-ORIGINAL benchmark performs better than our specialized variants of the FFORMA method for the Australian dataset but achieves substandard results for the New Zealand dataset. This suggests that we do not always need to create specialized ensembles for small areas, particularly when existing solutions have shown to work well for other demographic datasets (Makridakis et al., 2018; Montero-Manso et al., 2020).

The goal of this research was to investigate and develop competitive methods to improve the accuracy in of small area population forecasts. However, it is noteworthy to mention that there are many factors other than forecast accuracy that are considered by practitioners when choosing models for small area population forecasting. The availability of data is among such factors. Our demographic benchmark models only require data from the jump off year and 10 years prior, making them accessible to practitioners who only have access to census data. Practitioners often also need to explain why certain small areas are projected to increase whilst others are decreasing. Current machine learning models generally produce forecasts which are uninterpretable, however significant efforts are underway to address this and enable predictions made by machine learning models to be explained, and biases to be identified (Hardt et al., 2021). Other factors can include the computational time and the tools that are required to generate the forecasts. Machine learning methods are often computationally costly and consume significant amount of time to train and produce forecasts. For example, for the Australian dataset, the ESRNN model takes 1 hour and 50 minutes and the FFORMA-ORIGINAL model takes 50 minutes. However, LGBM is more computationally efficient, and our implementation only takes 5 minutes to train and generate forecasts for the Australian dataset. Conversely, the demographic models can be prepared with Microsoft Excel, making them readily accessible to

practitioners.

In this paper we investigated univariate time series forecasting methods for small areas. These methods are often appropriate for small area population forecasting as they require minimal data. However, it is important to note that there are other approaches that are not considered here, including cohort component models, housing-unit models, employment-led models, land-use/housing development models, regression models, microsimulation, and large-scale urban models (Wilson, 2011; Wilson et al., 2021). Combinations of these approaches have also been applied, such as the work of Cameron and Cochrane (2017) who use a land use model to disaggregate a regional-level cohort-component model to smaller areas. Future work should evaluate whether machine learning approaches can improve on the accuracy of these other population forecasting approaches. Additionally, there are other machine learning methods, such as Long Short-Term Memory (LSTM) networks that should also be further investigated in the global forecasting setting for the small area population context (Gers et al, 1999; Hochreiter & Schmidhuber, 1997). However, deep learning methods such as LSTM are more computationally costly and can take hours to run for datasets such as the Australian or New Zealand SA2 estimated resident populations. Conversely forecasts by LGBM models can be generated in minutes, which may make them a more appropriate option for practitioners.

5. Conclusions

Accurate small area population forecasts are very important for governments and businesses, so that they can anticipate the demand for public services in advance to optimize the allocation of resources. In this work, we have developed a framework to combine traditional demographic models and time series forecasting models for small area population forecasting. This includes an application of a machine learning based forecasting algorithm, LGBM, which has not been

evaluated for small area population forecasting to date. Our study also involves a comprehensive evaluation of forecast combination techniques used for aggregating population forecasts. We have tested this framework using two small area population datasets from Australia and New Zealand. The main findings of this research work are summarized below:

- Global LGBM models performed notably better than univariate LGBM models.
- The LGBM model is the best performing forecasting model for the New Zealand dataset.
- The FFORMA method is the best combination method for the New Zealand dataset.
- The demographic base models are the top performing individual models for the Australian dataset.
- The simple average-based combination methods, including the simple mean and the trimmed mean, produce the lowest errors overall for the Australian dataset.
- The proposed HORIZON-based combination method performs well across both of the datasets.
- LGBM is likely to perform better for the New Zealand dataset because its small areas population time series are more homogeneous than that of the Australian dataset.
- Combination methods do not always produce the most accuracy forecast, however they generally produce fewer relatively bad forecasts than the individual base models.
- The addition of LGBM to the ensemble models consistently improves the population forecast accuracy.

The results of this paper support the use of simple combination methods to improve forecast accuracy and decrease the prevalence of bad forecasts. Our findings also suggest that further research should be conducted to investigate whether the clustering of time series and application of localized models can help global methods, such as LGBM, to achieve more consistent forecast accuracy, particularly for small area population datasets that exhibit greater inter-time

series variability. We note that this work is specific for short forecast horizons. Future research is required to investigate whether similar results can be produced for longer forecast horizons.

This work demonstrates that state-of-the-art time series forecasting techniques can be used together with current demographic forecasting methods to improve the reliability of small area forecasts. We hope that future work in this research area can be helpful to promote and allow state of the art methods to be more widely available to practitioners.

6. Acknowledgements

The work is supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP200101480)

7. References

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, *16*, 521–530. https://doi.org/10.1016/S0169-2070(00)00066-2

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

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

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

Badmos, O.S., Rienow, A., Callo-Concha, D., Greve, K., & Juergens, C. (2019). Simulating slum growth in lagos: An integration of rule based and empirical based model. *Computers,*

environment and urban systems. 77, 101369. https://doi.org/10.1016/j.compenvurbsys.2019.101369

Bandara, K., Bergmeir, C., Campbell, S., Scott, D., & Lubman, D. (2020a). Towards accurate predictions and causal 'what-if' analyses for planning and policy-making: A case study in emergency medical services demand. *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10. https://doi.org/10.1109/IJCNN48605.2020.9206787

Bandara, K., Bergmeir, C., & Smyl, S. (2020b). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications, 140,* 112896. https://doi.org/10.1016/j.eswa.2019.112896

Bandara, K., Bergmeir, C., & Hewamalage, H. (2021). LSTM-MSNet: Leveraging forecasts on sets of related time series with multiple seasonal patterns. *IEEE Transactions on Neural Networks and Learning Systems, 32,* 1586–1599. https://doi.org/10.1109/TNNLS.2020.2985720

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 T, K. Wong, & M. Lee (Eds), *Neural Information Processing. ICONIP 2019. Lecture Notes in Computer Science* (vol 11955, pp. 462 - 474). Springer. https://doi.org/10.1007/978-3-030-36718-3_39

Bergstra, J., Yamins, D., & Cox, D.D. (2013). Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. *Proceedings of the 30th International Conference on International Conference on Machine Learning*, *28*(1), 115–123. https://proceedings.mlr.press/v28/bergstra13.html

Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning

opportunity. International Journal of Forecasting, 37(2), 587-603. https://doi.org/10.1016/j.ijforecast.2020.07.007

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Clemen, R.T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*, 559–583. https://doi.org/10.1016/0169-2070(89)90012-5

Diamond, I., Tesfaghiorghis, H., & Joshi, H. (1990). The uses and users of population projections in Australia. *Journal of the Australian Population Association 7*, 151–170. https://doi.org/10.1007/BF03029362

Eshragh, A., Ganim, B., Perkins, T., & Bandara, K. (2019). The importance of environmental factors in forecasting australian power demand, arXiv, 1911.00817. https://arxiv.org/abs/1911.00817

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information sciences, 180(10), 2044-2064. https://doi.org/10.1016/j.ins.2009.12.010

Genre, V., Kenny, G., Meyler, A., Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting 29*(1), 108–121. https://doi.org/10.1016/j.ijforecast.2012.06.004

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM. Paper presented at the 9th International Conference on Artificial Neural Networks: ICANN '99, Edinburgh, UK. https://doi.org/10.1049/cp:19991218

Godahewa, R., Bandara, K., Webb, G.I., Smyl, S., & Bergmeir, C. (2021). Ensembles of localised

models for time series forecasting. *Knowledge-Based Systems, 233*, 107518. https://doi.org/10.1016/j.knosys.2021.107518

Goodwin, P., (2009). New evidence on the value of combining forecasts. *Foresight: The International Journal of Applied Forecasting, 12*, 33–35.

Grossman, I., Wilson, T., & Temple, J. (2022, February 8). Forecasting small area populations with Long Short-Term Memory Networks. SocArXiv preprint. https://doi.org/10.31235/osf.io/3k79d

Hardt, M., Chen, X., Cheng, X., Donini, M., Gelman, J., Gollaprolu, S., . . . Kenthapadi, K. (2021). Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud. arXiv preprint. https://arxiv.org/abs/2109.03285

Hasegawa, Y., Sekimoto, Y., Seto, T., Fukushima, Y., & Maeda, M. (2019). My city forecast: Urban planning communication tool for citizen with national open data. *Computers, environment and urban systems 77*, 101255. http://dx.doi.org/10.1016/j.compenvurbsys.2018.06.001

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021a). Global models for time series forecasting: A simulation study. Pattern Recognition, 108441. https://doi.org/10.1016/j.patcog.2021.108441

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021b). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting 37*, 388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting, 21*, 15–24. https://doi.org/10.1016/j.ijforecast.2004.05.002

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hyndman, R., Wang, E., Kang, Y., Talagala, T., (2021a). *tsfeatures: Time Series Feature Extraction*. https://github.com/robjhyndman/tsfeatures/. r package version 0.1.

Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2021b). f*orecast: Forecasting functions for time series and linear models*. R package version 8.15. https://pkg.robjhyndman.com/forecast/

Hyndman, R.J., & Billah, B. (2003). Unmasking the theta method. *International Journal of Forecasting*, *19*, 287–290. https://doi.org/10.1016/S0169-2070(01)00143-1

Hyndman, R.J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *26*, 1–22. https://doi.org/10.18637/jss.v027.i03

Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., & Callot, L. (2020). Criteria for classifying forecasting methods. *International Journal of Forecasting*, *36*, 167–177. https://doi.org/10.1016/j.ijforecast.2019.05.008

Jolliffe I. (2011) Principal Component Analysis. In M. Lovric (Ed.). International Encyclopedia of Statistical Science (pp. 1094 - 1096). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_455

Kang, Y., Hyndman, R.J., & Li, F. (2020). GRATIS: GeneRAting TIme series with diverse and controllable characteristics. *Statistical Analysis and Data Mining. 13*(4), 354–376. https://doi.org/10.1002/sam.11461

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.Y., (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146 - 3154.Ke, G., Meng, Q., Finley, T., Wang, T.,

Chen, W., Ma, W., Ye, Q., & Liu, T.Y. (2021). Python Package. https://lightgbm.readthedocs.io/. Python Software Foundation.

Lichtendahl, K.C., & Winkler, R.L. (2020). Why do some combinations perform better than others? *International Journal of Forecasting, 36,* 142–149. https://doi.org/10.1016/j.ijforecast.2019.03.027

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting, 16,* 451–476. https://doi.org/10.1016/S0169-2070(00)00057-1

Makridakis, S., & Spiliotis, E. (2021). The M5 competition and the future of human expertise in forecasting. *Foresight: The International Journal of Applied Forecasting, 60,* 33–37.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting 34,* 802–808. https://doi.org/10.1016/j.ijforecast.2018.06.001

Makridakis, S., & Winkler, R.L. (1983). Averages of forecasts: Some empirical results. *Management Science, 29,* 987–996. https://doi.org/10.1287/mnsc.29.9.987

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. https://doi.org/10.1016/j.ijforecast.2019.02.011

Openshaw, S., & Van Der Knaap, G.A. (1983). Small area population forecasting: some experience with British models. *Journal of economic and social geography 74,* 291–304. https://doi.org/10.1111/j.1467-9663.1983.tb00976.x

Rayer, S. (2015). Demographic Techniques: Small-area Estimates and Projections. In International Encyclopedia of the Social & Behavioral Sciences (pp. 162-169).

https://doi.org/10.1016/B978-0-08-097086-8.31015-7

Rayer, S. (2007). Population forecast accuracy: Does the choice of summary measure of error matter? *Population research and policy review, 26,* 163–184. https://doi.org/10.1007/s11113-007-9030-0

Rayer, S., & Smith, S.K. (2010). Factors affecting the accuracy of subcounty population forecasts. *Journal of Planning Education and Research, 30*(2), 147–161. https://doi.org/10.1177%2F0739456X10380056

Rayer, S., Smith, S.K., & Tayman, J. (2009). Empirical prediction intervals for county population forecasts. *Population research and policy review, 28*, 773–793. https://doi.org/10.1007/s11113-009-9128-7

Riiman, V., Wilson, A., Milewicz, R., & Pirkelbauer, P. (2019). Comparing Artificial Neural Network and Cohort-Component Models for Population Forecasts. *Population Review*, 58(2). https://doi.org/10.1353/prv.2019.0008

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting 36,* 1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

Schapire, R.E. (1999). A brief introduction to boosting. P*roceedings of the 16th international joint conference on Artificial intelligence* (Volume 2, 1401-1406). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Shafizadeh-Moghadam, H. (2019). Improving spatial accuracy of urban growth simulation models using ensemble forecasting approaches. *Computers, environment and urban systems, 76*, 91–100. https://doi.org/10.1016/j.compenvurbsys.2019.04.005

Shaub, D., (2020). Fast and accurate yearly time series forecasting with forecast combinations.

International Journal of Forecasting 36, 116–120. https://doi.org/10.1016/j.ijforecast.2019.03.032

Shryock, H., & Siegel, J. (1973). *The Methods and Materials of Demography*. US Government Printing Office: Washington DC.

Smith, S.K., & Shahidullah, M. (1995). An evaluation of population projection errors for census tracts. *Journal of the American Statistical Association*, *90*, 64–71. https://doi.org/10.2307/2291130

Smith, S. K., Tayman, J., & Swanson, D. A. (2013). *A practitioner's guide to state and local population projections*: Springer. https://doi.org/10.1007/978-94-007-7551-0

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*, 75-85. https://doi.org/10.1016/j.ijforecast.2019.03.017

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

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

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

Timmermann, A. (2006). Chapter 4 forecast combinations, in: G. Elliott, C.W.J. Granger, & A. Timmermann (Eds). *Handbook of Economic Forecasting* (Vol. 1, pp. 135 - 196). Elsevier.

https://doi.org/10.1016/S1574-0706(05)01004-9

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. (2015). New evaluations of simple models for small area population forecasts. *Population, space and place, 21*, 335–353. https://doi.org/10.1002/psp.1847

Wilson, T., Brokensha, H., Rowe, F., & Simpson, L. (2018). Insights from the evaluation of past local area population forecasts. Population Research Policy Review, 37(1), 137–155. https://doi.org/10.1007/s11113-017-9450-4

Wilson, T., Grossman, I., Alexander, M., Rees, P., & Temple, J. (2021a). Methods for small area population forecasts: State-of-the-Art and research needs. *Population research and policy review*, 1 - 34. https://doi.org/10.1007/s11113-021-09671-6

Wilson, T., Grossman, I., & Temple, J. (2021b). Evaluation of the best M4 competition methods for small area population forecasting. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2021.09.005

Yang, Y & Hyndman, R.J (2021). *Introduction to the tsfeatures package*. https://pkg.robjhyndman.com/tsfeatures/articles/tsfeatures.html

Appendix A: The APE Ranking violin plots

Australia dataset

Figure A1. The APE Ranks for the 1-year forecast horizon for the Australia SA2 dataset

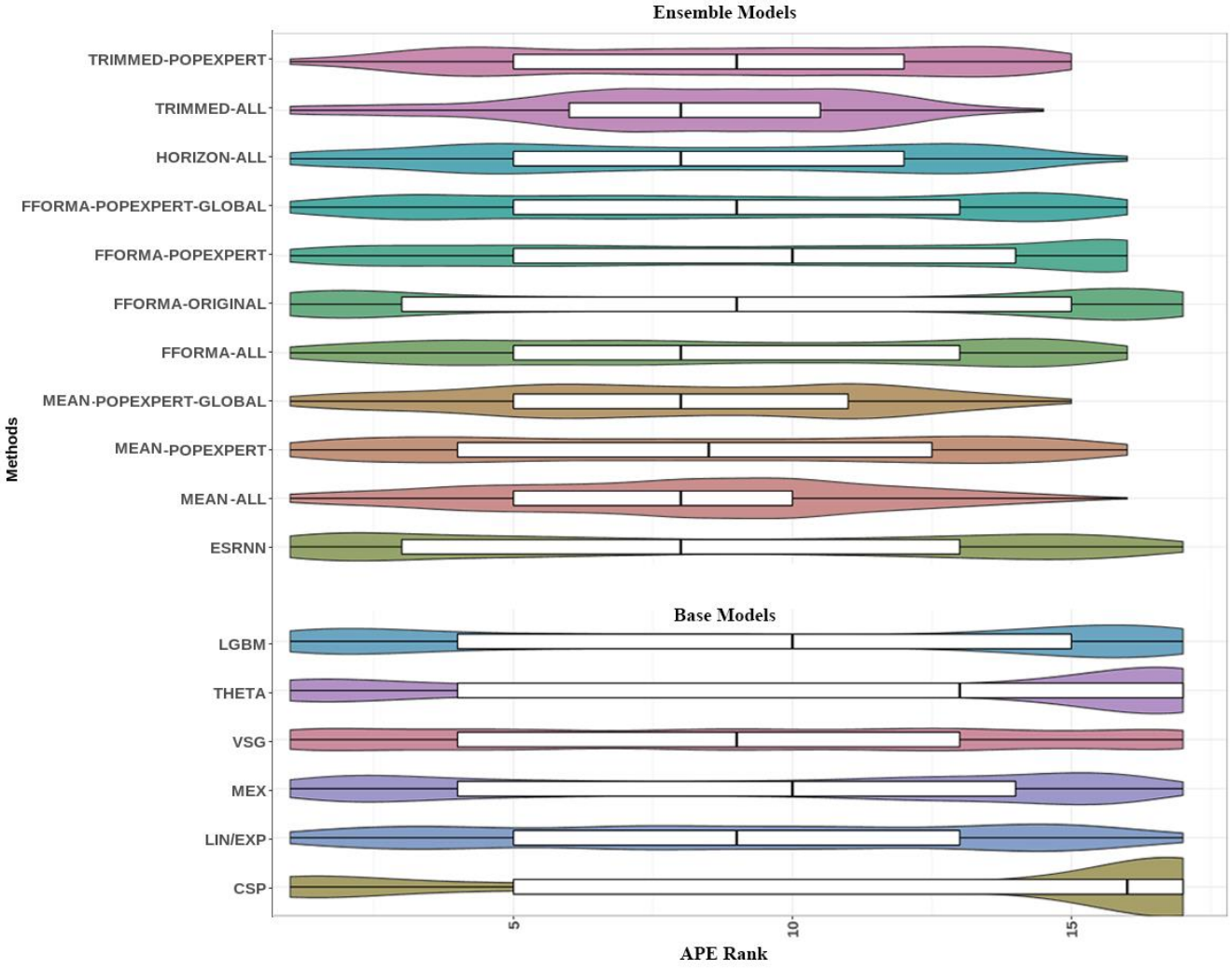


Figure A2. The APE Ranks for the 2-year forecast horizon for the Australia SA2 dataset

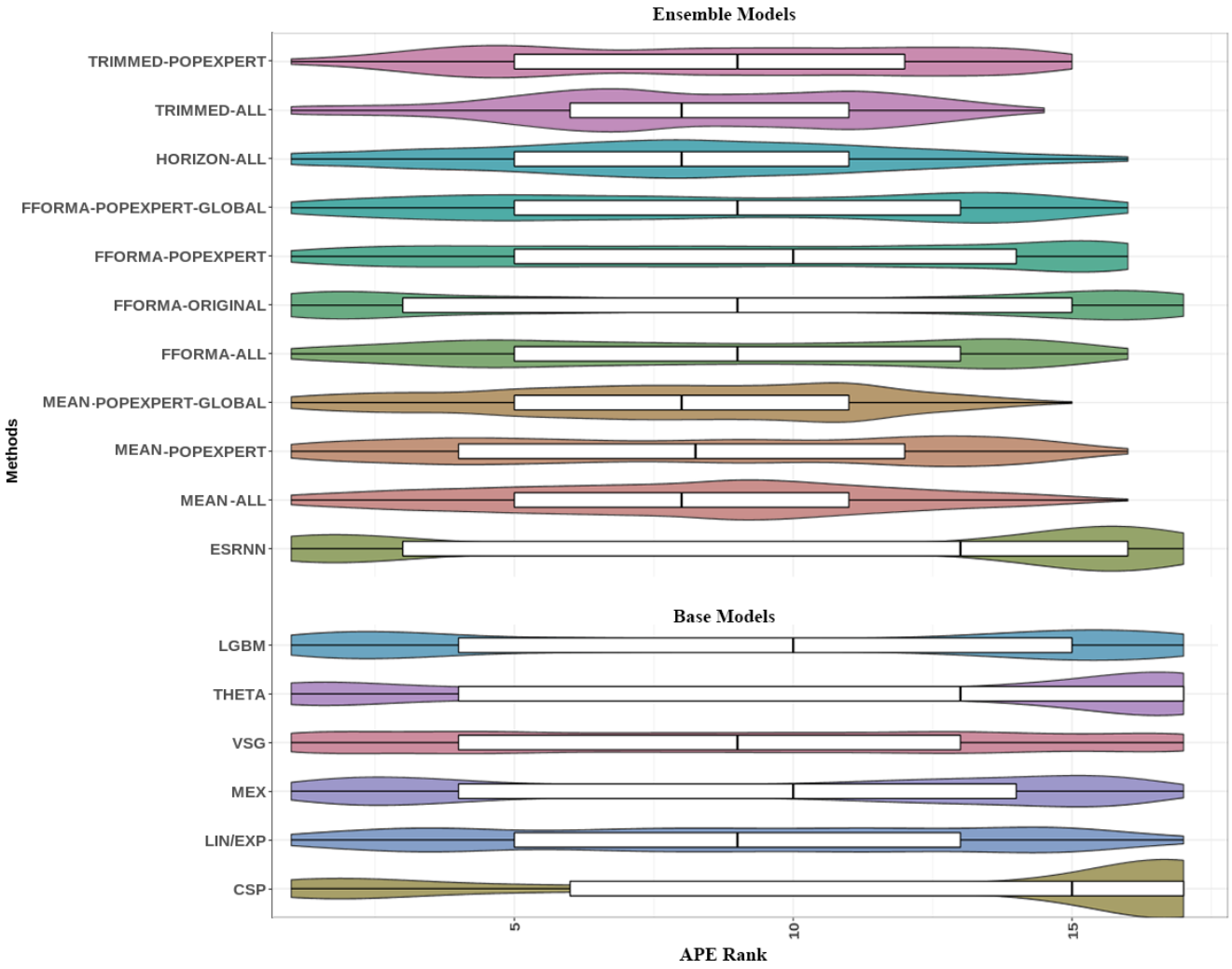


Figure A3. The APE Ranks for the 3-year forecast horizon for the Australia SA2 dataset

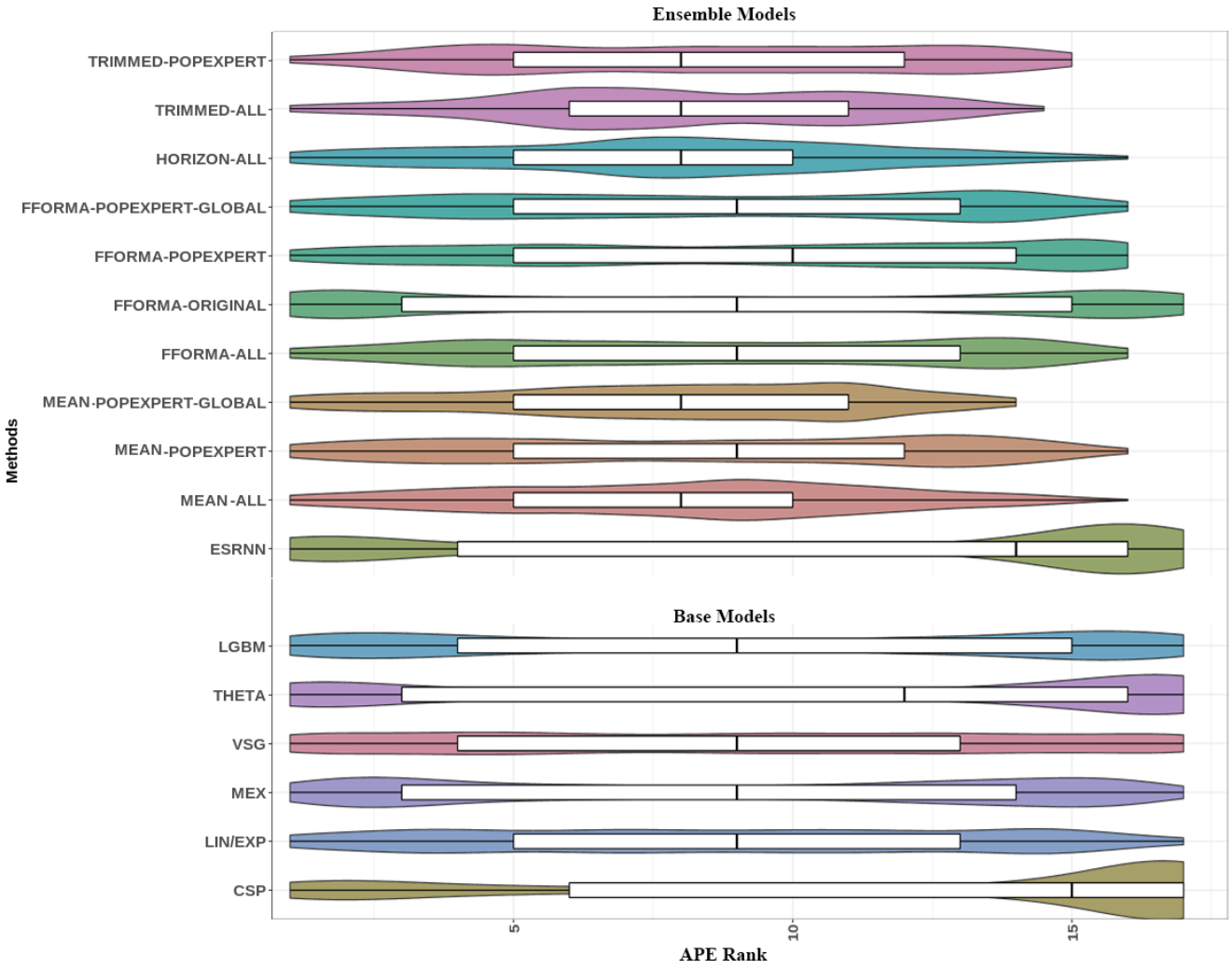
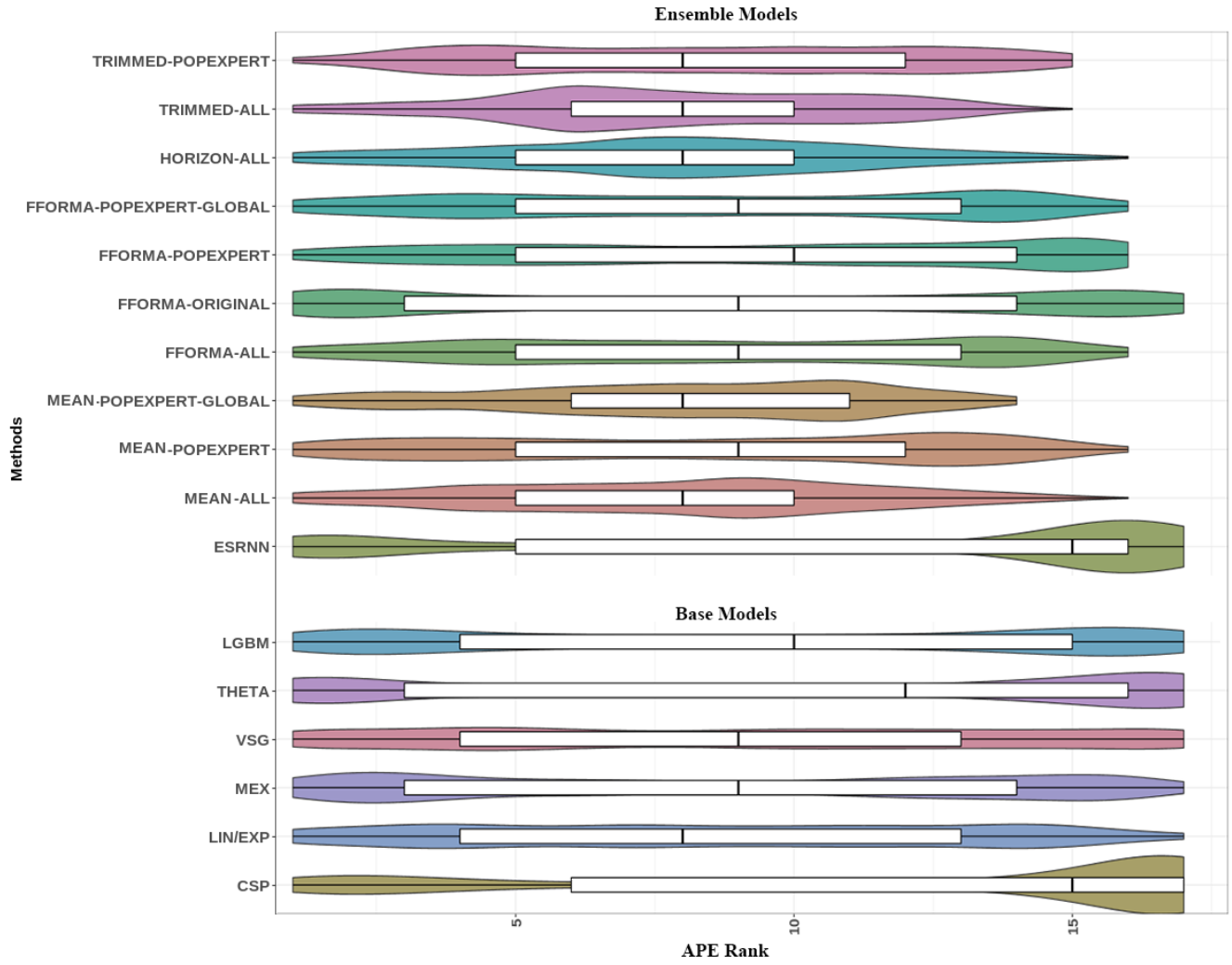


Figure A4. The APE Ranks for the 4-year forecast horizon for the Australia SA2 dataset



New Zealand dataset

Figure A5. The APE Ranks for the 1-year forecast horizon for the New Zealand SA2 dataset

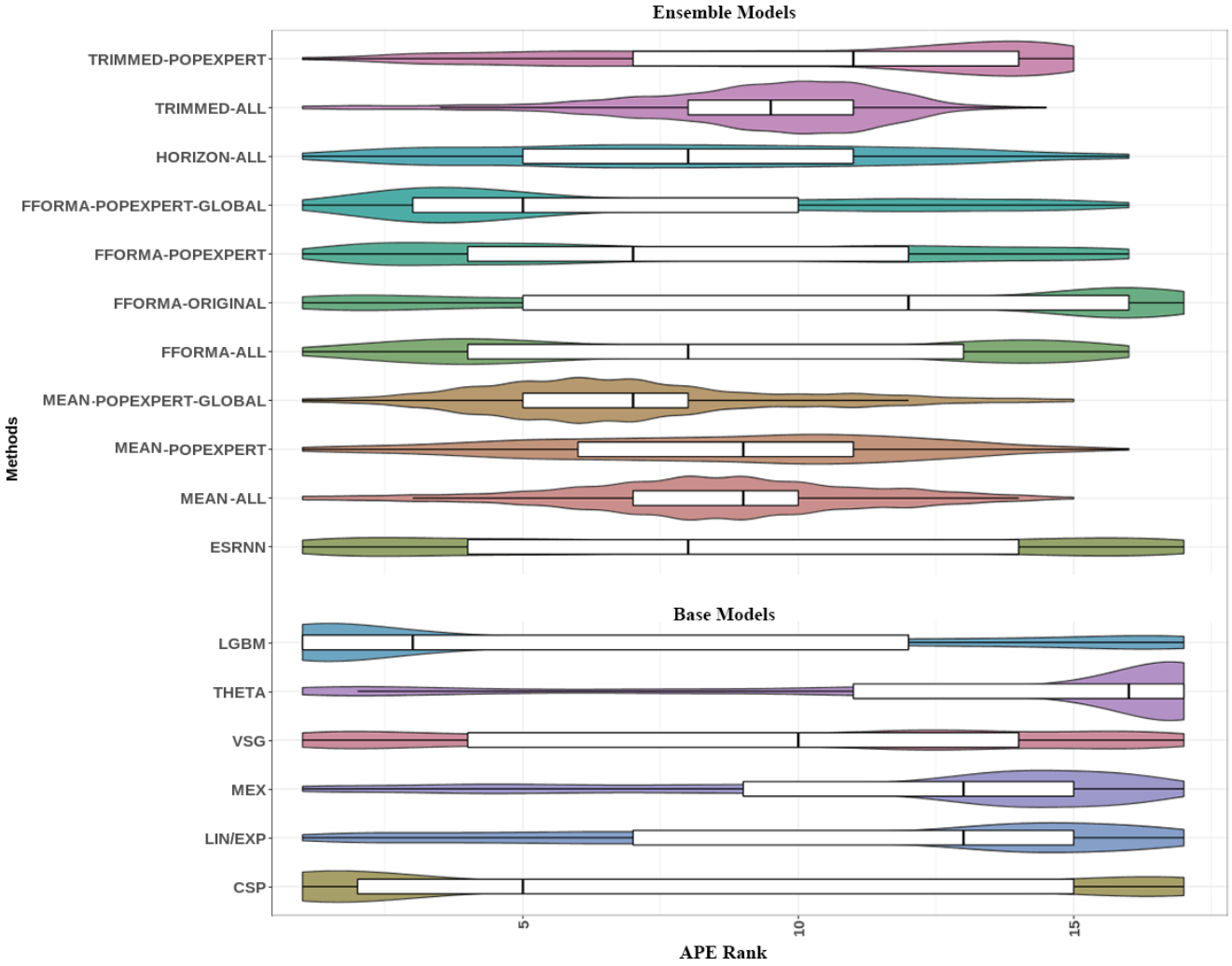


Figure A6. The APE Ranks for the 2-year forecast horizon for the New Zealand SA2 dataset

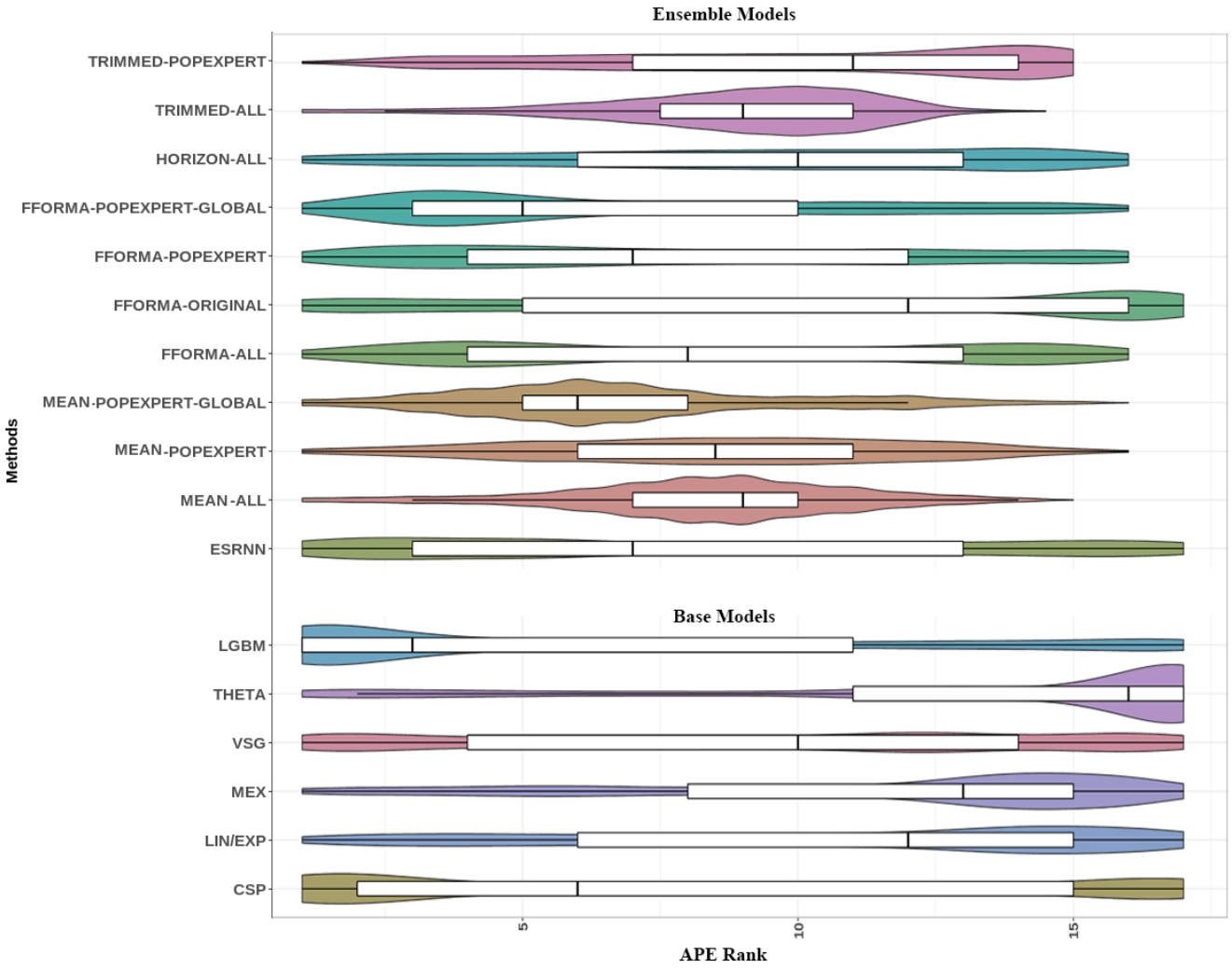


Figure A7. The APE Ranks for the 3-year forecast horizon for the New Zealand SA2 dataset

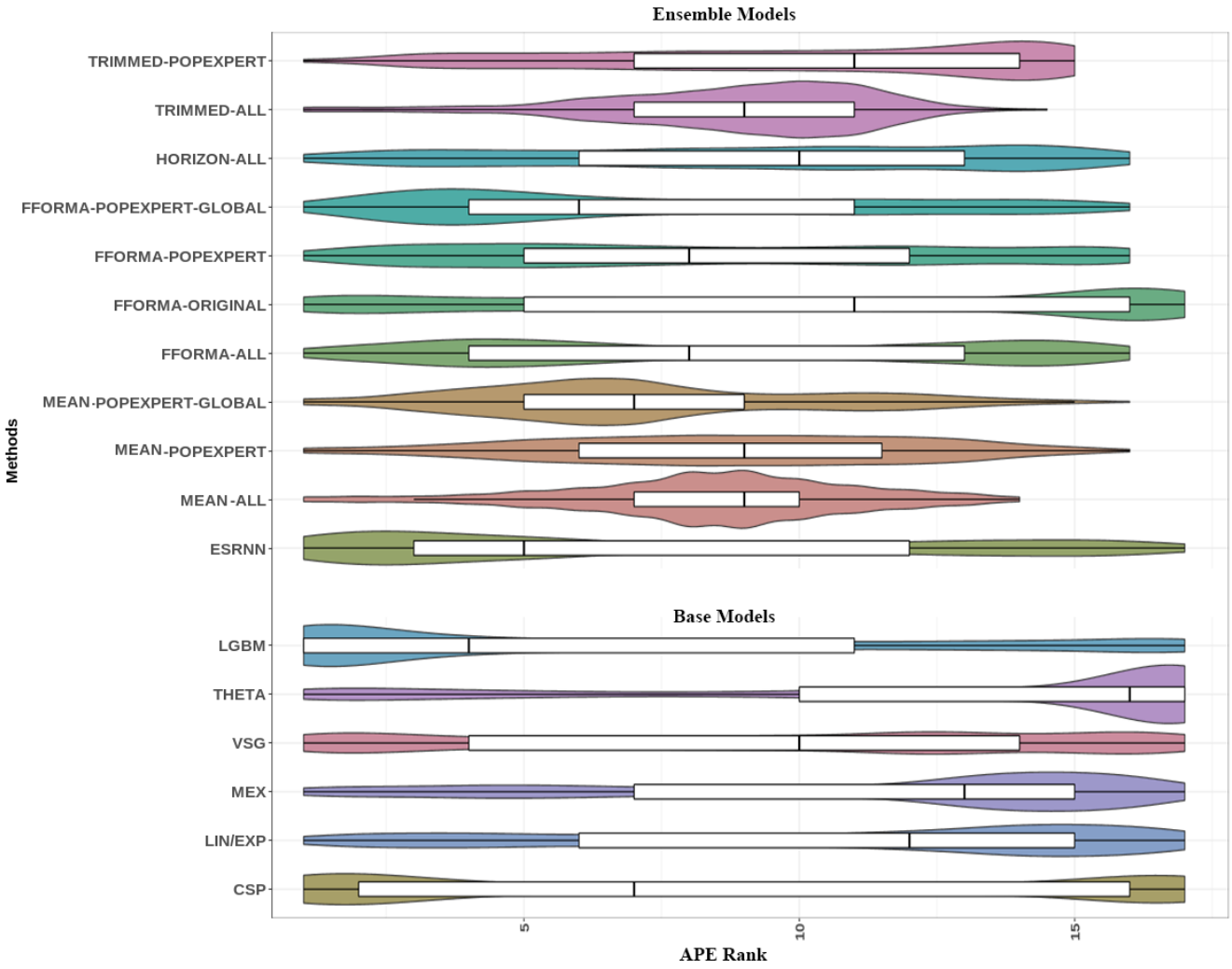
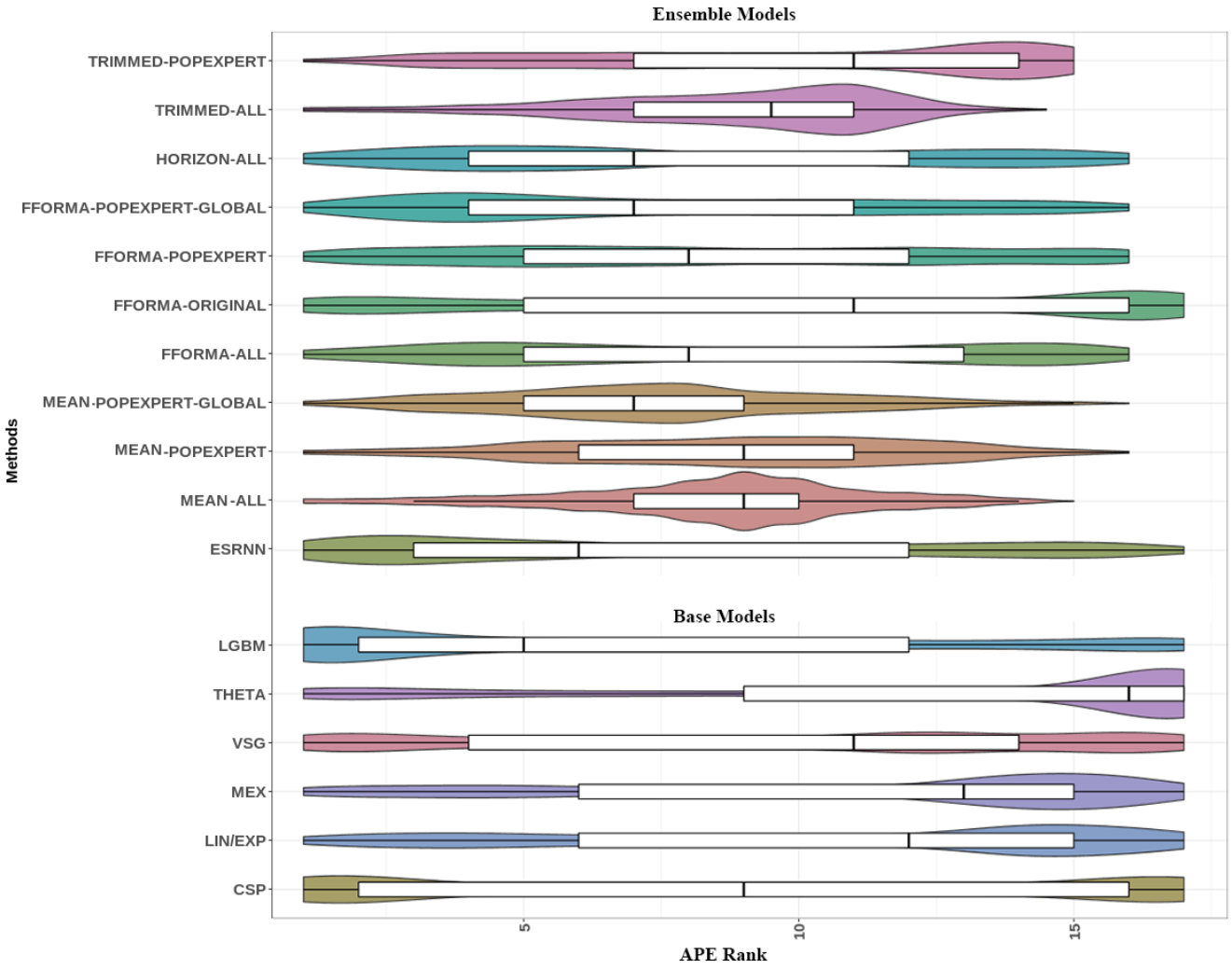


Figure A8. The APE Ranks for the 4-year forecast horizon for the New Zealand SA2 dataset



Appendix B: Significance Tests

Table B1: The Friedman rank sum test for the APE error measure for the 1st forecast horizon of the Australian dataset

Method	P_{Hoch}
MEAN-POPEXPERT-GLOBAL	-
MEAN-ALL	0.392
TRIMMED-ALL	0.158
HORIZON-ALL	6.678×10^{-2}
ESRNN	3.071×10^{-3}
MEAN-POPEXPERT	2.172×10^{-3}
FFORMA-ALL	1.924×10^{-6}
TRIMMED-POPEXPERT	1.924×10^{-6}
FFORMA-POPEXPERT-GLOBAL	8.008×10^{-7}
VSG	2.965×10^{-10}
LIN/EXP	7.435×10^{-13}
MEX	9.350×10^{-16}
FFORMA-ORIGINAL	1.413×10^{-16}
FFROMA-POPEXPERT	3.920×10^{-22}
LGBM	3.802×10^{-23}
THETA	1.488×10^{-73}
CSP	3.355×10^{-134}

Table B2: The Friedman rank sum test for the APE error measure for the 2nd forecast horizon of the Australian dataset

Method	P_{Hoch}
MEAN-POPEXPERT-GLOBAL	-
HORIZON-ALL	0.148
MEAN-ALL	0.148
TRIMMED-ALL	9.997×10^{-3}
MEAN-POPEXPERT	2.043×10^{-3}
TRIMMED-POPEXPERT	4.295×10^{-7}
FFORMA-POPEXPERT-GLOBAL	1.244×10^{-8}
FFORMA-ALL	3.621×10^{-9}
VSG	5.600×10^{-11}
LIN/EXP	2.532×10^{-14}
FFORMA-ORIGINAL	1.089×10^{-16}
MEX	6.104×10^{-19}
LGBM	2.964×10^{-23}
FFROMA-POPEXPERT	4.022×10^{-24}
ESRNN	1.968×10^{-48}
THETA	4.856×10^{-72}
CSP	1.684×10^{-150}

Table B3: The Friedman rank sum test for the APE error measure for the 3rd forecast horizon of the Australian dataset

Method	P_{Hoch}
MEAN-POPEXPERT-GLOBAL	-
HORIZON-ALL	0.659
MEAN-ALL	0.659
TRIMMED-ALL	0.623
MEAN-POPEXPERT	8.732×10^{-4}
TRIMMED-POPEXPERT	2.449×10^{-4}
FFORMA-POPEXPERT-GLOBAL	1.171×10^{-7}
LIN/EXP	1.820×10^{-8}
FFORMA-ALL	8.830×10^{-9}
MEX	4.588×10^{-10}
VSG	7.966×10^{-11}
FFORMA-ORIGINAL	2.381×10^{-12}
LGBM	1.775×10^{-19}
FFROMA-POPEXPERT	2.112×10^{-26}
THETA	1.008×10^{-53}
ESRNN	9.463×10^{-74}
CSP	2.547×10^{-140}

Table B4: The Friedman rank sum test for the APE error measure for the 4th forecast horizon of the Australian dataset

Method	P_{Hoch}
MEAN-ALL	-
TRIMMED-ALL	0.964
HORIZON-ALL	0.964
MEAN-POPEXPERT-GLOBAL	0.671
TRIMMED-POPEXPERT	5.969×10^{-4}
LIN/EXP	1.293×10^{-5}
MEAN-POPEXPERT	1.265×10^{-5}
MEX	2.395×10^{-6}
FFORMA-POPEXPERT-GLOBAL	1.678×10^{-7}
FFORMA-ALL	1.136×10^{-8}
FFORMA-ORIGINAL	1.249×10^{-10}
VSG	8.879×10^{-14}
LGBM	1.626×10^{-21}
FFROMA-POPEXPERT	6.738×10^{-30}
THETA	3.073×10^{-40}
ESRNN	6.335×10^{-99}
CSP	1.804×10^{-162}

Table B5: The Friedman rank sum test for the APE error measure for the 1st forecast horizon of the New Zealand dataset

Method	P_{Hoch}
LGBM	-
FFORMA-POPEXPERT-GLOBAL	7.240×10^{-3}
MEAN-POPEXPERT-GLOBAL	3.416×10^{-3}
CSP	4.124×10^{-16}
FFROMA-POPEXPERT	3.868×10^{-21}
HORIZON-ALL	3.569×10^{-23}
MEAN-ALL	1.342×10^{-36}
FFORMA-ALL	1.309×10^{-37}
MEAN-POPEXPERT	7.230×10^{-43}
ESRNN	4.649×10^{-52}
TRIMMED-ALL	4.808×10^{-66}
VSG	5.614×10^{-74}
TRIMMED-POPEXPERT	2.422×10^{-124}
FFORMA-ORIGINAL	1.672×10^{-135}
LIN/EXP	1.731×10^{-175}
MEX	3.258×10^{-238}
THETA	0

Table B6: The Friedman rank sum test for the APE error measure for the 2nd forecast horizon of the New Zealand dataset

Method	P_{Hoch}
LGBM	-
MEAN-POPEXPERT-GLOBAL	5.224×10^{-5}
FFORMA-POPEXPERT-GLOBAL	2.383×10^{-5}
FFROMA-POPEXPERT	4.747×10^{-28}
CSP	2.574×10^{-32}
ESRNN	4.808×10^{-39}
MEAN-ALL	8.381×10^{-41}
FFORMA-ALL	3.744×10^{-45}
MEAN-POPEXPERT	1.067×10^{-48}
TRIMMED-ALL	1.279×10^{-64}
VSG	3.477×10^{-89}
HORIZON-ALL	1.377×10^{-93}
TRIMMED-POPEXPERT	5.226×10^{-133}
FFORMA-ORIGINAL	7.855×10^{-151}
LIN/EXP	1.968×10^{-179}
MEX	5.231×10^{-245}
THETA	0

Table B7: The Friedman rank sum test for the APE error measure for the 3rd forecast horizon of the New Zealand dataset

Method	P_{Hoch}
LGBM	-
MEAN-POPEXPERT-GLOBAL	2.204×10^{-5}
ESRNN	1.220×10^{-6}
FFORMA-POPEXPERT-GLOBAL	8.268×10^{-7}
MEAN-ALL	1.571×10^{-33}
FFROMA-POPEXPERT	1.049×10^{-33}
FFORMA-ALL	1.157×10^{-39}
CSP	4.583×10^{-44}
MEAN-POPEXPERT	5.541×10^{-46}
TRIMMED-ALL	2.620×10^{-49}
HORIZON-ALL	2.513×10^{-72}
VSG	1.477×10^{-80}
TRIMMED-POPEXPERT	1.435×10^{-113}
FFORMA-ORIGINAL	2.452×10^{-116}
LIN/EXP	9.637×10^{-154}
MEX	1.182×10^{-192}
THETA	0

Table B8: The Friedman rank sum test for the APE error measure for the 4th forecast horizon of the New Zealand dataset

Method	P_{Hoch}
LGBM	-
MEAN-POPEXPERT-GLOBAL	3.018×10^{-5}
ESRNN	3.598×10^{-6}
FFORMA-POPEXPERT-GLOBAL	3.598×10^{-6}
HORIZON-ALL	5.642×10^{-16}
FFROMA-POPEXPERT	4.875×10^{-28}
MEAN-ALL	1.402×10^{-28}
FFORMA-ALL	4.669×10^{-34}
MEAN-POPEXPERT	1.024×10^{-38}
TRIMMED-ALL	9.719×10^{-40}
CSP	2.704×10^{-42}
VSG	8.078×10^{-70}
TRIMMED-POPEXPERT	1.503×10^{-93}
FFORMA-ORIGINAL	7.243×10^{-100}
LIN/EXP	2.804×10^{-128}
MEX	2.475×10^{-157}
THETA	2.383×10^{-290}