Hierarchical Time Series Forecasting in Emergency Medical Services

Abstract

Accurate forecasts of ambulance demand are crucial inputs when planning and deploying staff and fleet. Such demand forecasts are required at national, regional, and sub-regional levels, and must take account of the nature of incidents and their priorities. These forecasts are often generated independently by different teams within the organization. As a result, forecasts at different levels may be inconsistent, resulting in conflicting decisions and a lack of coherent coordination in the service. To address this issue, we exploit the hierarchical and grouped structure of the demand time series and apply forecast reconciliation methods to generate both point and probabilistic forecasts that are coherent and use all the available data at all levels of disaggregation. The methods are applied to daily incident data from an ambulance service in Great Britain, from October 2015 to July 2019, disaggregated by nature of incident, priority, managing health board, and control area. We use an ensemble of forecasting models and show that the resulting forecasts are better than any individual forecasting model. We validate the forecasting approach using time series cross validation.

Keywords: healthcare, emergency services, forecast reconciliation, ambulance demand, regression

1. Introduction

A failure to match available resources to demand in Emergency Medical Services (EMS) results in patient flow problems, with serious consequences for patients, staff, and the entire care system (Ekström et al., 2015; Rostami-Tabar and Ziel, 2022). Demand forecasting in EMS helps service planners to avoid the mismatch, potentially providing massive savings in costs and lives, and leading to better patient outcomes. Accurate daily demand forecasting enables planners and decision-makers to manage resources to meet anticipated patients, reconfigure units, and redeploy staff and vehicles as necessary.

Demand forecasts at EMS are typically required at multiple levels of an organization to inform various planning and decision-making processes (Hulshof et al., 2012). There are some planning processes at the national level (strategic and long-term) such as workforce resource planning and budgeting; sub-national, regional, or healthcare level (tactical and medium-term) such as temporary capacity expansions, resource sharing; and hospital or station level (operational and short-term) such as planning rosters for staff and ambulance deployment. Demand forecasts might also be required at different levels for a specific area of interest such as the nature of demand or the priority level. Moreover, the time series data in EMS has an inherent hierarchical and grouped structure to support such forecasting requirements. Demand for emergency medical services at the national level can be disaggregated in a geographical hierarchy into sub-national, regions, health boards, and stations/hospitals, or divided into groups such as the nature of incidents or demand priority. Forecasts produced at both higher and lower levels of hierarchies are necessary for effective decision-making in EMS. For example, control area EMS forecasts can inform strategic decisions about how to allocate limited resources to lower levels, such as health boards and stations/hospitals. At the lower levels, hospitals or ambulance stations could use such forecasts to plan for staffing and resource allocation, ambulance dispatching, staff-to-shift assignment, staff rescheduling based on the anticipated

*Corresponding author

volume and priority and nature of incidents. Additionally, generating forecasts at lower levels could potentially improve the accuracy of the high-level forecasts, by providing more detailed information on the nature and priority of incidents. This could help to identify patterns in demand that may not be apparent at the higher level. Therefore, employing forecasting techniques that consider the hierarchical and/or grouped patterns of time series in EMS aligns naturally, offering the possibility to enhance forecast accuracy and facilitate coordination.

However, despite a large number of studies dedicated to forecasting for EMS (Shi et al., 2022; Gul and Celik, 2020; Ibrahim et al., 2016; Wargon et al., 2009), the hierarchical data structure has been largely ignored, and the main focus has been on producing independent (base) forecasts at a single level. Generating independent forecasts can result in a lack of consistency and coordination, and therefore leads to less effective planning and decision making. For example, suppose the annual budget for the whole organization is to be allocated to each area health board based on demand within the corresponding area. If the forecasts are incoherent, there is a mismatch between the total resources to be allocated, and the resources allocated to each area. Further, suppose the budget also needs to take account of the nature of incidents that occur within each area, with more money allocated for some types of incidents than others. Then we need forecasts of the demand disaggregated by nature of incidents and health board, but these data are often too noisy to forecast directly. Forecast reconciliation addresses these problems by ensuring the forecasts are coherent across all aggregated and disaggregated series (avoiding mismatches), and by using the signal in the aggregated data to allow forecasting of highly disaggregated data (allowing better targetting of the budget).

With hierarchical forecasting, plans at any level are based on coherent forecasts and therefore can be aligned. Implementing and sustaining improvements in EMS require alignments and coordination between different stakeholders, without which teams operate in isolation leading to conflicts, duplication work, rework, or work that runs counter to the overall goal to improve the quality of delivery service. Hierarchical forecasting framework can be used as a tool to improve coordination between teams across the care services at the national, sub-national, regional and local levels. The hierarchical forecasting approaches not only create consistent forecasts but are usually also more accurate than the independent (base) forecasts (Hyndman et al., 2011). To our knowledge, there has been no previous research involving hierarchical and grouped forecasting in the entire field of forecasting for healthcare management.

In this paper, we address this gap by investigating the application of hierarchical forecasting approaches in the EMS using daily time series of attended incidents from 2015 to 2019 in a major ambulance service in Great Britain. The data has hierarchical and grouped structures, with hierarchies at the national, control (i.e. sub-national), and health board (i.e. regional) levels, as well as groups by priority and nature of incidents. We produce consistent point forecasts and forecast distributions for all levels, which is critical for an effective planning and associated risk management. We compare the point and probabilistic forecast accuracy of the independent forecasts, bottom-up and optimal reconciliation approaches. We first generate independent/base forecasts using Exponential Smoothing State Space (ETS), Poisson regression using Generalized Linear Model (GLM) and tscount (TSGLM), a simple empirical distribution and an ensemble method, followed by applying bottom-up and optimal reconciliation approaches. Forecast performance is assessed by the Mean Absolute Scaled Error (MASE) and Mean Squared Scaled Error (MSSE) for point forecasts and Continuous Ranked Probability Scores (CRPS) for the probabilistic forecasts. This paper complies with reproducibility principles (Stodden and Miguez; Boylan et al., 2015). We provide the R codes for the proposed models and benchmarks. Therefore, they can be applied to any healthcare service (e.g., emergency department, primary or social care) subject to the time series having a hierarchical and/or grouped structure.

The remainder of this article is structured as follows: In Section 2, we provide a brief review of the literature and discuss its limitation to position our work; in Section 3, we present the experiment design describing the data set, forecasting methods and forecast evaluation metrics. In Section 4, we discuss the hierarchical time series forecasting approaches to generate both point and probabilistic forecasts. In Section 5, we present and discuss our results; in Section 6, we summarize our findings and present ideas for future research.

2. Research background

Emergency medical services (EMS) are a critical component in the delivery of urgent medical care to communities. An effective service delivery requires accurate resource planning that generally relies on demand forecasts at operational, tactical, and strategic levels. There is a substantial number of studies on the application of time series forecasting in the Emergency Medical Services. For example, Ibrahim et al. (2016) provide an extensive review of the models used in forecasting call volume arrivals. Another important area is related to forecasting ambulance demand. Although the definition of demand might not be always clearly stated, this is typically referring to a situation where a physical resource has been deployed to respond to an incident. This might be also called *attended incidents*. Another demand related variable is verified incidents; these are all incidents that require an action: either by sending a physical vehicle, responding via the Clinical Support Desk, requesting an external provider to respond to it, or forwarding it to other channels such as police, firefighters, or general practitioners. Our study is aligned with this stream of literature. Another similar area that has received considerable attention is Emergency Department forecasting; we refer interested readers to Shi et al. (2022), Gul and Celik (2020), and Wargon et al. (2009) for extensive reviews of the relevant literature. In this section, we provide a brief review of studies on forecasting ambulance demand in EMS.

There are generally two main streams of research related to forecasting ambulance demand in EMS: (i) the first stream focuses on the application of time series methods and regression approaches to forecasting aggregate ambulance demand (Vile et al., 2012; Sasaki et al., 2010); and (ii) the second stream considers forecasting EMS demand in finer temporal and geographical granularities by employing temporal-spatial prediction methods (Zhou and Matteson, 2016; Zhou, 2016). The focus of our study is related to the first stream of research.

Sasaki et al. (2010) develop a multivariable regression model to estimate future EMS demands. In addition to the historical demand, the population census for different age groups and counts of the number of companies employing more than five people are included in the regression. The census variables describe groups who are more likely to need an ambulance. A stepwise ordinary least squares regression analysis is used for estimating the parameter and generating forecasts. The only performance measure reported in this study is R^2, which is not an effective measure of forecast accuracy (Armstrong, 2001, p457). The research design of this study is not rigorous and the study is not reproducible. Vile et al. (2012) explore using a Singular Spectrum Analysis (SSA) method to generate forecasts of the EMS demand at the national level for 7-day, 14-day, 21-day, and 28-day forecast horizons using data provided by an ambulance service in Great Britain. The performance of this approach is compared to Auto-Regressive Integrated Moving Average (ARIMA) and Holt-Winters time series methods using Root Mean Squared Error (RMSE). They concluded that point forecasts generated by SSA are more accurate for longer-term, but that ARIMA and Holt-Winters performance is superior for shorter-term horizons. Vile et al. (2016) further develop a decision support system to integrate forecasts generated by SSA. However, the study does not compare and contrast the performance of forecasting methods based on utility measures such as cost, resource utilization and response time. The tool contains options that allow generating forecasts at various levels of granularity; however, it ignores the structure of the hierarchical and grouped relationships, preventing aligned decision making and coordination. Al-Azzani et al. (2021) utilises data from the Welsh Ambulance Service to explore the forecast accuracy of four forecasting approaches: ARIMA, Holt Winters, Multiple Regression, and Singular Spectrum Analysis (SSA) in predicting call volume demand. The aim is to compare these approaches with the current method across various planning horizons (7 days, 30 days, and 90 days) for both total call volume and category-specific demand. Forecast accuracy performance is evaluated using root mean square error (RMSE) and mean absolute percentage error (MAPE). The findings indicate that ARIMA performs the best in predicting weekly and monthly demand. However, when it comes to long-term demand, the SSA method proves to be the most effective. Ibrahim et al. (2016) conducted a case study to assess the effectiveness of multiple forecasting methods: the multiplicative univariate forecasting model (MU), univariate mixed-effects model (ME), bivariate mixed-effects model (BME1), and bivariate mixed-effects model (BME2). Call centre data were utilised to forecast for periods of 1, 7, and 14 days ahead, using only a limited dataset of 42 days.

The performance of these forecasting methods was evaluated using two metrics: RMSE for point forecasts and coverage probability for the 95% prediction interval. The findings indicate that the ME consistently produces the most accurate point forecasts. On the other hand, BME1 and BME2 demonstrate superior coverage probabilities when forecasting for one day or one week ahead. For a two-week leading period, MU shows better coverage probability.

Hermansen and Mengshoel (2021) investigate forecasting EMS demand in a high Spatio-temporal resolution of 1km^2 spatial regions and 1-hr time intervals using total incidents in Oslo, Norway, from 1 January 2015 to 11 February 2019. They used multi-layer perceptron (MLP) and long short-term memory (LSTM) models to forecast the EMS demand, and compare the results to simple aggregation methods and baselines. The point forecast accuracy is evaluated using Mean Absolute Error (MAE) and Mean Squared Error (MSE), and the forecast distribution is measured by Categorical Cross-Entropy. They found that Neural Network models performed better in producing point forecasts, while a distribution baseline method based on the spatial distribution of the incidents across all time steps provided more accurate forecast distributions. Zhou (2016) proposed three methods based on Gaussian mixture models, kernel density estimation, and kernel warping to predict hourly data 4 weeks ahead for a 1km^2 spatial region. Two years of incidents from Toronto, Canada (years 2007 and 2008 with 391,296 events) and Melbourne, Australia were used to build the model and examine the performance on test data using mean negative log-likelihood. They show that forecasts generated by the proposed methods were significantly more accurate than the current industry practice (a simple averaging formula). Grekousis and Liu (2019) investigated the combination of spatial analysis methods with data mining techniques based on an improved Hungarian algorithm and a MLP neural network to identify the most likely locations of future emergency events. The proposed approach was tested using data from 2851 events attended by the EMS in Athens, Greece, over 24 weeks. They showed that 23% of real emergency events lie within 50 meters of the predicted ones and nearly 70% of the real emergency events lie no further than 150 meters away, which is rather accurate given the granularity of the problem at the city level.

We note a number of limitations in the literature of EMS forecasting, that encourage us to undertake this research. These limitations are summarized as following:

1. Current studies ignore the inherent hierarchical and/or grouped structure of the time series data, and the relationship between series at different levels of hierarchy. While the hierarchical forecasting methodology has been developed and applied in various domains over the past 10 years (**?**), it has never been explored in this area.
2. Current research is mainly concerned with generating point forecasts at a single level of hierarchy. There is a lack of studies considering the entire forecast distribution of daily ambulance demand for the whole hierarchy to better represent the uncertainty of future demand, providing a risk management tool for planners.
3. Reproducibility is still a major challenge in EMS forecasting, as it is unlikely that any reader can reproduce prior studies without the help of the authors of those papers.
4. Another limitation is related to the generated forecasts not being on the sample space of non-negative counts. Since actual ambulance counts cannot be negative or non-integer, ambulance demand forecast distributions should reflect the data. Of course, point forecasts represent means, so they should be non-negative, but may be non-integer. While this might not be an issue when producing forecasts at a single level, producing non-negative count forecasts in a hierarchical/grouped structure is challenging and requires further investigation in the future.

Table 1: Summary of some studies in forecasting FOR AMBULANCE SERVICES

Reference	Year	Variable	Horizon	Method	Metric	Probabilistic	Reconciliation
Current study	2023	Ambulance demand	84 days	Naïve; Exponential Smoothing State Space (ETS), Poisson regression using Generalized Linear Model (GLM) and tscount (TSGLM), a simple empirical distribution and an ensemble method	MASE, MSSE, CRPS	YES	YES
Al-Azzani et al.	2021	Call volume	7 , 30 , 90 days	ARIMA, Holt Winters, Multiple Regression, and Singular Spectrum Analysis	RMSE, MAPE	NO	NO
Haugsbø et al.	2021	Ambulance demand in Spatio-temporal	1hour	MLP, LSTM	MAE, MSE, Cross-Entropy	YES	NO
Grekousis et al.	2019	Locations of incidents	1 hour	MLP and Hungarian algorithm	RMSE	NO	NO
Ibrahim et al.	2016	Call volume	1, 7 , 14 days	multiplicative univariate forecasting, univariate mixed-effects, bivariate mixed-effects model, and bivariate mixed-effects	RMSE, prediction interval coverage	Partial	NO
Vile et al.	2012	Ambulance demand	7, 14, 21, 28 day	Singular Spectrum Analysis, ARIMA, Holt-Winters	RMSE	NO	NO
Sasaki et al.	2010	Ambulance demand	5 years	OLS regression	R^2	NO	NO

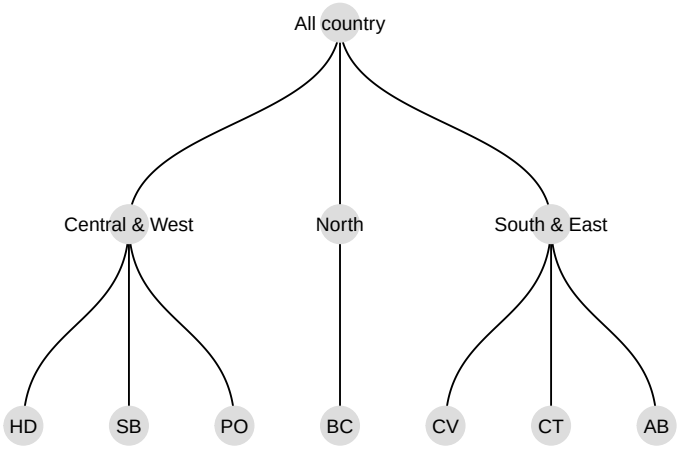
3. Experiment setup

Planners in the ambulance service work with a planning horizon of 6 weeks. That is, planning is generally frozen for the next 42 days, so any forecasts will only affect plans for the time period beyond the next 42 days. Consequently, the forecast horizon in this study is $2 \times 42 = 84$ days ahead, with performance evaluation assessed based on the last 42 days and not the whole forecast period. The forecasts are produced for various training and test sets using time series cross-validation (Hyndman and Athanasopoulos, 2021).

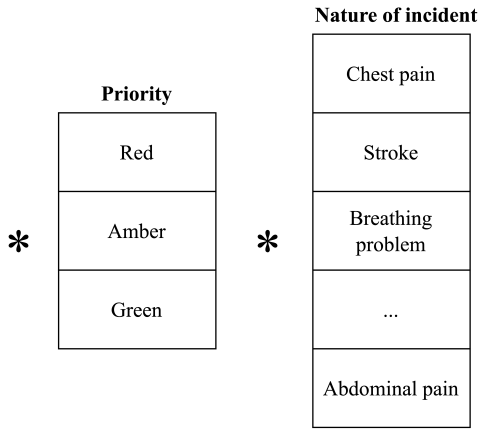
In the following section, we discuss the dataset, describe the forecasting methods used to generate base forecasts, and present the point and probabilistic accuracy measures.

3.1. Data

The dataset used in this study is from a major ambulance service in Great Britain. It contains information relating to the daily number of attended incidents from 1 October 2015 to 31 July 2019, disaggregated by nature of incidents, priority, the health board managing the service and the control area (or region). Figure 1 depicts both the hierarchical and grouped structure of the data. Figure 1a illustrates the nested hierarchical structure based on control area and health board and Figure 1b shows the grouped structure by priority and the nature of incident.



(a) Hierarchical structure: Attended incidents in the whole country are disaggregated into 3 control areas and then into 7 different healthboards.



(b) Grouped structure: Incidents could be grouped into priority (i.e. Red, Amber & Green) and the nature of attended incident (i.e. there are 35 different nature of incidents including chest pain, breathing problems, heart attack, stroke, and so on). The symbol * refers to the crossed attributes between hierarchical and grouped levels.

Figure 1: The hierarchical and grouped structure of attended incidents (ambulance demand).

Table 2 also displays the structure of data with the total number of series at each level. At the top level, we have the total attended incidents for the country. We can split these total attended incidents by control area, by health board, by priority or by nature of incident. There are 3 control areas breakdown by 7 local health boards. Attended incident data are categorized into 3 priority classes of red, amber, and green. There are also 35 different nature of incidents such as chest pain, stroke, breathing problem, etc. In total, across all levels of disaggregation, there are 1530 time series.

Given the total number of time series, direct visual analysis is infeasible. Therefore, we first compute features of all 1530 time series (Kang et al., 2017) and display the strength of trend and weekly seasonality strength in Figure 2. Each point represents one time series with the strength of trend in x-axis and the strength of seasonality in y-axis. Both measures are on a scale of [0,1].

Table 2: Number of time series in each level for the hierarchical & grouped structure of attended incidents

Level	Number of series
All country	1
Control	3
Health board	7
Priority	3
Priority * Control	9
Priority * Health board	21
Nature of incident	35
Nature of incident * Control	105
Nature of incident * Health board	245
Priority * Nature of incident	104
Control * Priority * Nature of incident	306
Control * Health board * Priority * Nature of incident (Bottom level)	691
Total	1530

Note:
Due to certain combinations of the nature of incident with other variables, there is a lack of representation in the dataset. As a result, for example, instead of the calculation 3 * 35 = 105, it would be modified to 3 * 35-1 = 104.

In this paper, the strength of trend and seasonality were calculated using the "STL" (Seasonal and Trend decomposition using Loess) decomposition method, as described by Bandara et al. (in press). STL is a widely used and flexible method for decomposing time series data into trend, seasonal, and remainder components. The decomposition of a time series y_t is written as $y_t = T_t + S_t + R_t$, where T_t is the smoothed trend component, S_t is the seasonal component and R_t is a remainder component. The strength of trend is defined as:

$$F_T = \max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(T_t + R_t)}\right)$$

For strongly trended data, the seasonally adjusted data should have much more variation than the remainder component. Therefore $\text{Var}(R_t)/\text{Var}(T_t + R_t)$ should be relatively small. But for data with little or no trend, the two variances should be approximately the same.

The strength of seasonality is defined similarly:

$$F_S = \max\left(0, 1 - \frac{\text{Var}(R_t)}{\text{Var}(S_t + R_t)}\right).$$

series with seasonal strength F_S, close to 0 exhibits almost no seasonality, while a series with strong seasonality will have F_S close to 1 because $\text{Var}(R_t)$ will be much smaller than $\text{Var}(S_t + R_t)$.

It is clear that there are some series showing strong trends and/or seasonality, corresponding to series at the higher levels of the hierarchy. The majority of series show low trend and seasonality. These are time series belonging to the bottom series, series related to the nature of incidents for a given control, health board and priority level. Bottom series are dominated by noise with little or no systematic patterns.

In addition to displaying the trend and seasonality features, we also visualize a few time series at various levels of aggregation. Figure 3 reveals different information such as trend, seasonality, and noise. For example, some series depict seasonality and trend, whereas some other series report low volume of attended incidents and entropy, making them more volatile and difficult to forecast. At the level on nature of incidents combined with categories of other levels, there are many series that contain zeros with low counts. As such, the data set represents a diverse set of daily time series patterns.

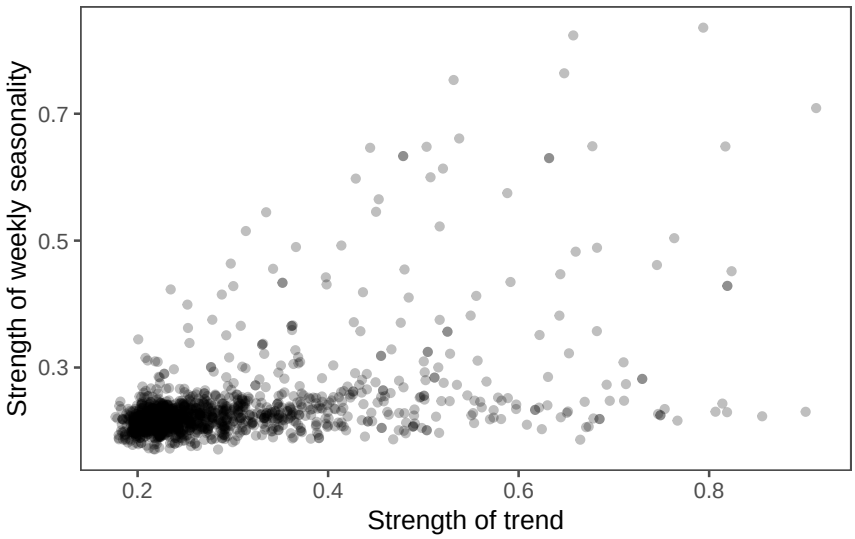


Figure 2: The strength of the trend and weekly seasonality in the time series of attended incidents. The scatter plot shows a total of 1530 data points, with each point corresponding to a specific time series.

We consider several forecasting models that account for the diverse patterns of the time series across the entire hierarchy. In developing the forecasting models, the time series of holidays are also used in addition to the attended incidents. We use public holidays, school holidays and Christmas Day and New Year's Day as predictors of incident attended. These types of holidays will affect peoples' activities and may increase or decrease the number of attended incidents.

3.2. Forecasting methods

Given the presence of various patterns in the past attended incidents, we consider three different forecasting models to generate the base forecasts. Once the base forecasts are produced, hierarchical and grouped time series methods are used to reconcile them across all levels. We briefly discuss forecasting models in the following sections, and the hierarchical forecasting methods are discussed in Section 4.

Stationary: We start with a simple forecasting approach, assuming that the future days will be similar to past days. We use the empirical distribution of the past daily attended incidents to create the forecast distribution of future attended incidents. We have chosen this "stationary" method as a benchmark due to its widespread usage and simplicity, making it easily understandable for users. Forecasts serve as inputs for various decision-making systems that frequently employ simulations, wherein it is common to utilize the empirical distribution of demand as a forecast. Additionally, the stationary method has shown surprisingly high accuracy. Hence, any forecasting approach that can offer superior results compared to the stationary method would validate its practical use, otherwise there is no necessity for employing more complex methods.

Exponential Smoothing State Space model (ETS): ETS models (Hyndman and Athanasopoulos, 2021) can combine trend, seasonality, and error components in a time series through various forms that can be additive, multiplicative or mixed. The trend component can be none ("N"), Additive ("A") or damped ("Ad"); the seasonality can be none ("N"), Additive ("A"), or multiplicative ("M"); and the error term can be additive ("A") or multiplicative ("M"). To forecast the attended incidents at each level, we use the `ets()` function in the `forecast` package (Hyndman et al., 2022; Hyndman and Khandakar, 2008) in R. To identify the best model for a given time series, the `ets` function uses the corrected Akaike's Information Criterion (AICc).

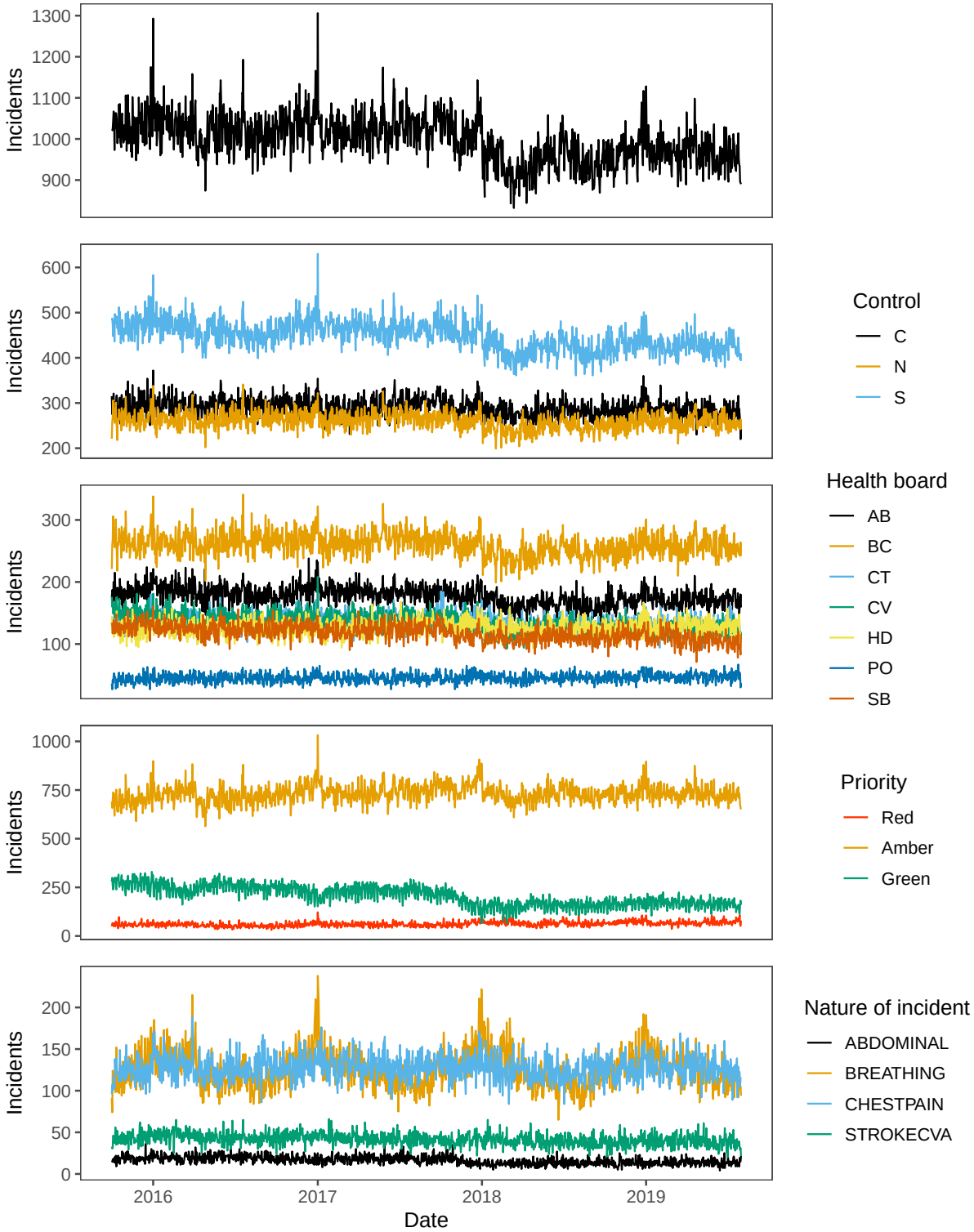


Figure 3: Daily time plot of attended incidents at various levels. X-axis shows the date of incidents, consisting of 1400 data points (days) and y-axis shows the number of attended incidents. The panels show data from the whole country (top panel), by control area, by health board, by priority level, and by nature of incident. Only four of the 35 nature of incident categories are shown to avoid too much overplotting. Each time plot .

In our study, we use an automated algorithm to determine the suitable configuration for the trend, seasonality, and error terms in each time series. Specifically, we utilize the 'ETS()' function in the fable package of R, which employs Akaike's Information Criterion (AIC) to identify the optimal model for each time series. Given the large number of time series we work with (1530), it is impractical to manually select the appropriate form for each component in every time series. Consequently, the automated algorithm selects the best model based on the unique characteristics of each individual time series. As a result, a combination of additive or multiplicative forms for the components are employed, depending on the specific attributes of each time series.

Despite the popularity and the relevance of automatic `ETS` in this study, it may produce forecast distributions that are non-integer and include negative values, although the number of attended incidents is always integer and non-negative. When using `ETS`, a time series transformation approach could be used to generate strictly positive forecasts, although forecast distributions will still be non-integer. An alternative is to use forecasting models that produce integer, non-negative forecasts. In the following section we present Generalized Linear Models (GLMs) and Poisson time series regression to produce count base forecasts.

Generalized Linear Model (GLM): GLMs are a family of models developed to extend the concept of linear regression models to non-Gaussian distributions (Faraway, 2016). They model the response variable as a particular member of the exponential family, with the mean being a transformation of a linear function of the predictors. One of the models that is frequently used in practice to generate count forecasts is Poisson regression.

Suppose the time series is denoted by $y_1, \ldots, y_T$, then the Poisson GLM can be written as

$$\begin{aligned} & y_t \sim \text{Poisson}(\lambda_t) \\ \text{where} \qquad & \log(\lambda_t) = \boldsymbol{x}_t'\boldsymbol{\beta}, \end{aligned}$$

and $\boldsymbol{x}_t$ is a vector of covariates, β is a vector of coefficients, and λ_t is the mean of the Poisson distribution. In our model, these include cubic splines for the time trend, day-of-week dummy variables (from Monday to Sunday), Fourier terms to capture the yearly seasonality, dummy variables indicating public holidays (1 when is public holiday, 0 otherwise), school holidays (1 when is school holiday, 0 otherwise) and Christmas Day (1 when is Christmas Day, 0 otherwise) and New Year's Day (1 when is New Year's Day, 0 otherwise). The Fourier terms are as defined in Hyndman and Athanasopoulos (2021, Section 7.4) This model takes account of weekly seasonality and annual seasonality. Monthly seasonality is exceedingly rare in time series data, and does not occur in ambulance demand. There is no reason, for example, for incidents to occur more at some times of the month than others.

We fit a Poisson regression model using the function `glm()` from the *stats* package in R, with the argument `family = poisson` to specify that we wish to fit a Poisson regression model with a log link function.

Poisson Regression using tscount (TSGLM): We also consider another Poisson regression model that takes into account serial dependence. This model captures the short range serial dependence by including autoregressive terms, in addition to the same covariates that were used in the GLM model. To distinguish this from the previous GLM model, we will refer to this model as `TSGLM`.

The Poisson TSGLM is similar to GLM with an additional autoregressive component accounting for serial correlation. The term serial dependence refers to situations where the ambulance demand at a given time point exhibits correlation with the demand at previous time points, after accounting for the exogenous covariates. Put simply, past values of the time series influence the current values, and various techniques can be employed to model this influence. TSGLM can be written as

$$\begin{aligned} & y_t \sim \text{Poisson}(\lambda_t) \\ \text{where} \qquad & \log(\lambda_t) = (\boldsymbol{y}_{t-k}', \boldsymbol{x}_t')\boldsymbol{\beta}, \end{aligned}$$

and y_{t-k} is a vector of k lagged values. The TSGLM model explicitly accounts for serial dependence by including lagged values (i.e. past values) of the ambulance demand in the model. This is important in EMS

forecasting because it allows the model to capture patterns in the data that are dependent on the past values of the time series, which might not be captured via the predictor variables.

We use the 'tsglm()' function in the 'tscount' package in R (Liboschik et al., 2017) to model the attended incidents. Again, the logarithmic link function is used to ensure that the mean of the Poisson distribution is always positive.

Provided accidents occur independently, they will inherently follow a Poisson distribution (Feller, 1991, p156–158). Hence, it is reasonable to assume a Poisson distribution in this context. To account for changes over time, we incorporate trend and seasonality covariates, as well as public holiday effects, allowing the mean of the Poisson distribution to vary. However, it is important to note that if there are additional factors influencing the mean of the Poisson distribution, which are not accounted for in our model, we might observe over-dispersion or under-dispersion in the data.

Ensemble method: Finally, one effective strategy for improving forecast accuracy includes the simultaneous application of multiple forecasting methods on a given time series, followed by combining the forecasts, rather than relying on separate forecasts generated by each individual method (Clemen, 1989). In this paper, we use an ensemble method that combines the forecasts generated from the Stationary, ETS, GLM and TSGLM models using a simple average to form a mixture distribution (Wang et al., in press).

To generate forecast probability distributions using above methods, we use a form of bootstrapping, described in Panagiotelis et al. (2023). This involves simulating 1000 future sample paths from each of the models, by bootstrapping the model residuals, taking into account the cross-sectional correlations between the different aggregated and disaggregated series. In this way, we can generate an empirical distribution of forecasts for each model. The ensemble forecast distribution is a simple mixture of these empirical distributions.

It is important to emphasize that the aim of this study is not to provide an exhaustive compilation of forecasting models, or to promote a particular model class. Instead, we have developed a flexible framework that can accommodate any forecasting models. Our primary objective is to demonstrate its practicality and effectiveness in integrating base forecasts from any model and generating coherent forecasts within a hierarchical structure.

3.3. Performance evaluation

To evaluate the performance of the various forecasting approaches, we split the data into a series of ten training and test sets. We use a time series cross-validation approach (Hyndman and Athanasopoulos, 2021), with a forecast horizon of 84 days, and each training set expanding in 42-day steps. The first training set uses all data up to 2018-04-25, and the first test set uses the 84 days beginning 2018-04-26. The second training set uses all data up to 2018-06-06, with the second test set using the following 84 days. The largest training set ends on 2019-05-09, with the test set ending on 2019-07-31. Model development and hyper-parameter tuning is performed using the training data and the errors are assess using the corresponding test set. While we compute forecast errors for the entire 12 weeks, we are most interested in the last 42 days of each test set, because that corresponds to how forecasts are generated for planning in practice. Forecasting performance is evaluated using both point and probabilistic error measures.

The error metrics provided below consider a forecasting horizon denoted by j, representing the number of time periods ahead we are predicting. In our study, this forecasting horizon ranges from 1 to 84 days, $j = 1, 2, \ldots, 84$.

Point forecast accuracy is measured via the Mean Squared Scaled Error (MSSE) and the Mean Absolute Scaled Error (MASE). The Mean Absolute Scaled Error (MASE) (Hyndman and Koehler, 2006; Hyndman and Athanasopoulos, 2021) is calculated as:

$$\text{MASE} = \text{mean}(|q_j|),$$

where

$$q_j = \frac{e_j}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|},$$

and e_j is the point forecast error for forecast horizon j, $m = 7$ (as we have daily seasonal series), y_t is the observation for period t, and T is the sample size (the number of observations used for training the forecasting model). The denominator is the mean absolute error of the seasonal naive method in the fitting sample of T observations and is used to scale the error. Smaller MASE values suggest more accurate forecasts. Note that the measure is scale-independent, thus allowing us to average the results across series.

A related measure is MSSE (Hyndman and Athanasopoulos, 2021; Makridakis et al., 2022), which uses squared errors rather than absolute errors:

$$\text{MSSE} = \text{mean}(q_j^2),$$

where,

$$q_j^2 = \frac{e_j^2}{\frac{1}{T-m}\sum_{t=m+1}^{T}(y_t - y_{t-m})^2},$$

Again, this is scale-independent, and smaller MSSE values suggest more accurate forecasts.

Using scale-independent measures, such as MASE and MSSE, enables more appropriate comparisons between time series at different levels and scales, as these measures are not influenced by the magnitude of the data. This is of particular importance in our study, as we work with time series at various levels of hierarchy, with varying scales, resulting in different magnitudes of error. By employing scale-independent measures, we can meaningfully assess the forecast accuracy across the entire hierarchy, ensuring a more robust comparison.

To measure the forecast distribution accuracy, we calculate the Continuous Rank Probability Score (Gneiting and Katzfuss, 2014; Hyndman and Athanasopoulos, 2021). It rewards sharpness and penalizes miscalibration, so it measures overall performance of the forecast distribution.

$$\text{CRPS} = \text{mean}(p_j),$$

where

$$p_j = \int_{-\infty}^{\infty} \left(G_j(x) - F_j(x)\right)^2 dx,$$

where $G_j(x)$ is the forecasted probability distribution function for forecast horizon j, and $F_j(x)$ is the true probability distribution function for the same period.

Calibration refers to the statistical consistency between the distributional forecasts and the observations. It measures how well the predicted probabilities match the observations. On the other hand, sharpness refers to the concentration of the forecast distributions — a sharp forecast distribution results in narrow prediction intervals, indicating high confidence in the forecast. A model is well-calibrated if the predicted probabilities match the distribution of the observations, and it is sharp if it is confident in its predictions. The CRPS rewards sharpness and calibration by assigning lower scores to forecasts with sharper distributions, and to forecasts that are well-calibrated. Thus, it is a metric that combines both sharpness and miscalibration into a single score, making it a useful tool for evaluating the performance of probabilistic forecasts.

CRPS can be considered an average of all possible Winkler scores (Winkler, 1972; Hyndman and Athanasopoulos, 2021, Section 5.9) or percentile scores (Hyndman and Athanasopoulos, 2021, Section 5.9), and thus provides an evaluation of all possible prediction intervals or quantiles. A specific prediction interval could be evaluated using a Winkler score. Certain situations may also require assessing accuracy for a particular quantile, such as lower (e.g 5%) or higher (e.g. 95%) quantiles. In such cases, a percentile score becomes useful in meeting this specific requirement.

4. Hierarchical and grouped time series forecasting techniques

There are many applications in healthcare, and in particular in EMS, where a collection of time series is available. These series are generally hierarchically organized based on multiple levels such as area/region, health board and/or are aggregated at different levels in groups based on nature of demand, priority of demand, or some other attributes. While series could be strictly hierarchical or only grouped bases on some attributes, in many situations a more complex structures arise when attributes of interest are both nested and crossed, having hierarchical and grouped structure. This is also the case for our application as discussed in Section 3.1.

4.1. Independent (base forecast)

A common practice in healthcare (and EMS) to predict hierarchical and grouped series relies on producing independent forecasts, also refereed to as base forecasts, typically by different teams as the need for such forecasts arise. We observe n time series at time t, across the entire hierarchical and grouped structure, written as y_t. The base forecasts of y_{T+h} given data $y_1, \ldots, y_T$ are denoted by $\hat{y}_h$ for h steps-ahead for all n series ($n = 1530$ in this study). Forecasts generated in this way are not coherent.

4.2. Reconciliation methods

Traditionally, approaches to produce coherent forecasts for hierarchical and grouped time series involve using bottom-up and top-down methods by generating forecasts at a single level and then aggregating or disaggregating. Top-down methods require having a unique hierarchical structure to disaggregate forecasts generated at the top level by proportions. However, given that we have multiple grouped attributes combined with the hierarchical structure, there is no unique way to disaggregate top forecasts. Hence the top-down approach cannot be used in our application. The recommended approach is to use forecast reconciliation (Hyndman et al., 2011). In the following sections, we first discuss some notation, and then present bottom-up and forecast reconciliation approaches used in this study to generate coherent forecasts.

4.2.1. Notations

Let $\boldsymbol{b}_t$ be a vector of n_b *bottom-level* time series at time t, and let $\boldsymbol{a}_t$ be a corresponding vector of $n_a = n - n_b$ aggregated time series, where

$$\boldsymbol{a}_t = \boldsymbol{A}\boldsymbol{b}_t,$$

and $\boldsymbol{A}$ is the $n_a \times n_b$ "aggregation" matrix specifying how the bottom-level series $\boldsymbol{b}_t$ are to be aggregated to form $\boldsymbol{a}_t$. The aggregation matrix $\boldsymbol{A}$ is determined by the structure of the hierarchy. It maps the bottom-level time series to the corresponding higher-level time series. For example, if there are two bottom-level series, and one aggregated series (equal to the sum of the two bottom-level series), then $\boldsymbol{A} = \begin{bmatrix} 1 & 1 \end{bmatrix}$. The full vector of time series is given by

$$\boldsymbol{y}_t = \begin{bmatrix} \boldsymbol{a}_t \\ \boldsymbol{b}_t \end{bmatrix}.$$

This leads to the $n \times n_b$ "summing" or "structural" matrix given by

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{I}_{n_b} \end{bmatrix}$$

such that $\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t$.

The term "bottom-level series" relates to the most disaggregated series within the hierarchical and grouped time series structure. For instance, in Table 2, each distinct combination of values in Control area (e.g. South & East), Health board (e.g. CV), Priority (e.g. Green), and Nature of incident (e.g. Chest pain), corresponds

to one individual time series. In the dataset at hand, there are 691 unique combinations, resulting in 691 bottom level time series. The "aggregate time series" describes how these bottom-level series are combined to create higher-level series. For instance, to obtain the incidents at the national level (i.e. all country level), the time series are aggregated across all Control areas, Health boards, Priorities, and Natures of incidents. Any desired aggregation level can be achieved based on the data structure, utilizing the bottom-level series available.

4.2.2. Bottom-up (BU) and linear reconciliation methods

Bottom-Up is a simple approach to generate coherent forecasts. It involves first creating the base forecasts for the bottom-level series (i.e., the most disaggregated series). These forecasts are then aggregated to the upper levels which naturally results in coherent forecasts. The BU approach can capture the dynamics of the series at the bottom level, but these series may be noisy and difficult to forecast. The approach uses only the data at the most disaggregated level, and so does not utilize all the information available across the hierarchical and grouped structure.

The bottom-up (BU) approach is constrained by its reliance solely on base forecasts from a single level of aggregation at the bottom level. While it does result in consistent forecasts, the BU approach lacks forecast reconciliation since no reconciliation is performed.

Forecast reconciliation approaches bridge this gap by combining and reconciling all base forecasts to generate coherent forecasts. This technique utilizes all the base forecasts produced within a hierarchical structure to create consistent forecasts at every level of the hierarchy. As a result, it goes beyond relying solely on base forecasts from a single level of aggregation, and instead leverages all available information at each level to generate forecasts that minimize the total forecast variance of the set of coherent forecasts. Linear reconciliation involves projecting the base forecasts onto the coherent space. It is derived by minimizing the sum of the variances of the reconciled forecasts subject to the resulting forecasts being coherent and unbiased (Wickramasuriya et al., 2019).

Linear reconciliation methods can be written (Wickramasuriya et al., 2019) as

$$\tilde{\boldsymbol{y}}_h = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}^{-1}\boldsymbol{S})^{-1}\boldsymbol{W}^{-1}\hat{\boldsymbol{y}}_h,$$

where $\boldsymbol{W}$ is an $n \times n$ positive definite matrix, and $\hat{\boldsymbol{y}}_h$ contains the h-step forecasts of $\boldsymbol{y}_{T+h}$ given data to time T. Different choices for $\boldsymbol{W}$ lead to different solutions such as Ordinary Least Squares (OLS), Weighted Least Squares (WLS) and Minimum Trace (MinT).

Ordinary Least Squares (OLS) is the simplest and most commonly used method for estimating the parameters in linear regression models. In this approach, the estimation of $\boldsymbol{W}$ is based on the assumption that all the errors have equal variance. Hence, $\boldsymbol{W}$ is simply defined as the identity matrix multiplied by a constant factor. The intuition behind OLS is that it minimizes the sum of squared residuals between the observed and predicted values of the dependent variable. The main weakness of this approach is that it does not take account of the different scales of the base time series; the aggregated series will usually have higher variance than the disaggregated series, simply because the values are larger, but OLS treats all series the same. A strength of the approach is that it is simple, and does not involve estimating a covariance matrix.

Weighted Least Squares (WLS) is an extension of OLS where the variance of the errors is assumed to be heteroscedastic, i.e., different for each series. But it assumes that the errors of each series are uncorrelated with each other. In this approach, $\boldsymbol{W}$ is defined as a diagonal matrix with the variance of the errors on the diagonal. The intuition behind WLS is that it assigns higher weight to series with smaller error variance, and thereby takes into account the different scales of the base time series. The main weakness of this approach is that it ignores the relationships between series. A strength of WLS is that it is relatively easy to compute $\boldsymbol{W}$ as it is based only on error variances which are readily estimated.

Minimum Trace (MinT) is a further generalization where $\boldsymbol{W}$ is defined as the covariance matrix of the base forecast errors. So it takes account of both the scale of each series, and the relationships between the series.

Wickramasuriya et al. (2019) showed that this approach gives the optimal reconciled forecasts in the sense that the sum of the forecast variances is minimized. The main weakness of this approach is that it is difficult to estimate the full covariance matrix. In practice, we usually need to use a shrinkage estimate where the off-diagonal elements are shrunk towards zero.

We use the implementation of these methods in the fable package in R in the experiment.

Certainly, other approaches can be applied to hierarchical forecasting problems Pennings and Van Dalen (2017) and Villegas and Pedregal (2018) proposed the idea of using a state space model to ensure consistent forecasts. However, when dealing with larger hierarchies, these models encounter difficulties in estimating covariance matrices. In contrast, our approach provides a clear advantage by allowing the incorporation of different forecasting methods for the base forecasts, and even accommodating distinct methods for individual series. The decoupling of time series models from the reconciliation step adds significant flexibility in exploring a wide range of models.

5. Results and discussion

In this section, we compare the forecasting performance of the Stationary, ETS, GLM, and TSGLM models along with the ensemble, using base forecast and Minimum Trace (MinT) reconciliation methods. We have also computed the forecast accuracy for Ordinary Least Square (OLS) and Weighted Least Square (WLS) approaches, along with bottom-up forecasting. However, they are not reported here because their accuracy is outperformed by MinT. We should also note that forecasts, and consequently their corresponding errors, are generated for the entire hierarchy and they could be reported at any level, if required. But to save space, we have reported only the top level (Total), the bottom level, and the levels corresponding to Control areas and Health boards. The latters are chosen because this is where decision-making takes place, so these forecasts are the most important.

The overall forecasting performance is reported in Table 3, in which the average forecast accuracy over horizons 43–84 days (corresponding to the planning horizon) is presented per model, method, and the hierarchical level. Reported forecast accuracy is averaged across all forecast horizons, rolling origins, and series at each level. Table 3 presents both point and probabilistic forecast accuracy at total, control area, health board and bottom-level series. Point forecast performance is reported using MASE and MSSE in Table 3a and 3b, respectively. Probabilistic forecast accuracy is reported using CRPS in Table 3c. The bold entries in each table identify a combination of method and model that performs best for the corresponding level (i.e. each column), based on the smallest values of accuracy measures.

Table 3a and 3b show that forecast reconciliation (i.e. MinT) improves forecast accuracy at the higher levels of the hierarchy including total, control area and health board. However, it does not result in accuracy improvement at the bottom-level series, for which base forecasts are more accurate. This might be due to the noisy structure of time series at the bottom level, and possibly due to very different patterns in the aggregated series. It is also clear from Table 3a that the ensemble method improves forecast accuracy at total, control area and health board. However, this does not remain valid for bottom series where different individual methods perform best, depending on the accuracy measure. While the forecast reconciliation approach aims to enhance forecast accuracy, its effectiveness is not guaranteed, especially if the bottom-level series exhibit excessive noise and lack systematic patterns. Despite this, reconciling forecasts at the bottom level can offer advantages by generating coherent forecasts that facilitate alignment in planning across various teams within an organization, promote better coordination, and prevent conflicting decisions. Moreover, even when dealing with noisy and irregular bottom-level series, reconciliation can still improve forecast accuracy at higher levels of the hierarchy by leveraging the information available across the hierarchy. Therefore, although the bottom-level forecasts may not be highly accurate on their own, reconciling them with higher-level forecasts can still provide a more consistent view of future demand and potentially yield more accurate forecasts at other levels.

Table 3: Average forecast performance calculated on the test sets at forecast horizons $h = 43, \ldots, 84$ days, with time series cross validation applied to attended incident data. The testset consists of 462 days. The best approach is highlighted in bold.

(a) Point forecast accuracy using MASE

		MASE			
Method	Model	Total	Control areas	Health boards	Bottom
Base	Stationary	1.139	1.059	1.047	1.019
Base	ETS	0.963	0.930	0.899	1.038
Base	GLM	0.910	0.940	0.923	**1.002**
Base	TSGLM	0.911	0.939	0.924	1.005
Base	Ensemble	0.782	0.856	0.876	1.008
MinT	Stationary	1.138	1.059	1.047	2.651
MinT	ETS	0.877	0.916	0.915	1.289
MinT	GLM	0.848	0.901	0.902	2.493
MinT	TSGLM	0.852	0.903	0.903	2.513
MinT	Ensemble	**0.753**	**0.844**	**0.872**	2.260

(b) Point forecast accuracy using MSSE

		MSSE			
Method	Model	Total	Control areas	Health boards	Bottom
Base	Stationary	1.169	1.056	1.062	1.031
Base	ETS	0.979	0.875	0.816	**0.975**
Base	GLM	0.813	0.897	0.875	1.009
Base	TSGLM	0.822	0.901	0.875	1.050
Base	Ensemble	0.599	0.729	0.774	0.993
MinT	Stationary	1.168	1.057	1.062	2.095
MinT	ETS	0.785	0.852	0.845	0.994
MinT	GLM	0.720	0.827	0.837	1.803
MinT	TSGLM	0.722	0.833	0.839	1.851
MinT	Ensemble	**0.560**	**0.706**	**0.765**	1.557

(c) Probabilistic forecast accuracy using CRPS

		CRPS			
Method	Model	Total	Control areas	Health boards	Bottom
Base	Stationary	30.387	10.882	5.500	0.302
Base	ETS	14.309	6.074	3.476	0.244
Base	GLM	15.396	6.253	3.576	0.244
Base	TSGLM	15.316	6.227	3.575	0.245
Base	Ensemble	12.978	**5.727**	3.430	0.243
MinT	Stationary	30.368	10.902	5.498	0.313
MinT	ETS	13.515	5.967	3.547	**0.243**
MinT	GLM	13.839	5.917	3.453	0.246
MinT	TSGLM	14.000	5.947	3.455	0.248
MinT	Ensemble	**12.585**	5.728	**3.426**	0.247

Table 3c presents the accuracy of the forecast distributions measures by CRPS, which considers both forecasting reliability and interval sharpness. The smaller the value of CRPS, the better the comprehensive performance. We observe that forecast reconciliation results in forecast improvement for the total and health board level. CRPS is almost identical at the control area and botom levels. Base forecasts are slightly better at the control area level, while reconciliation is marginally accurate than base at the bottom level. The ensemble method is also more accurate for higher levels, but ETS performs well at the bottom level. Table 3c indicates that reconciliation using Mint generates accurate *distributional* forecasts. The marginal improvement in the average probabilistic forecast accuracy at the bottom level might be due to the reconciliation method giving improved forecast accuracy in the tails of the forecast distribution, which are critical for managing risks.

Overall, our results indicate that forecast reconciliation using the MinT method provides reliable forecasts and improves upon the base (unreconciled) forecasts at all levels except the bottom-level series. But even there, forecast reconciliation using MinT improves accuracy in the tails of the distribution.

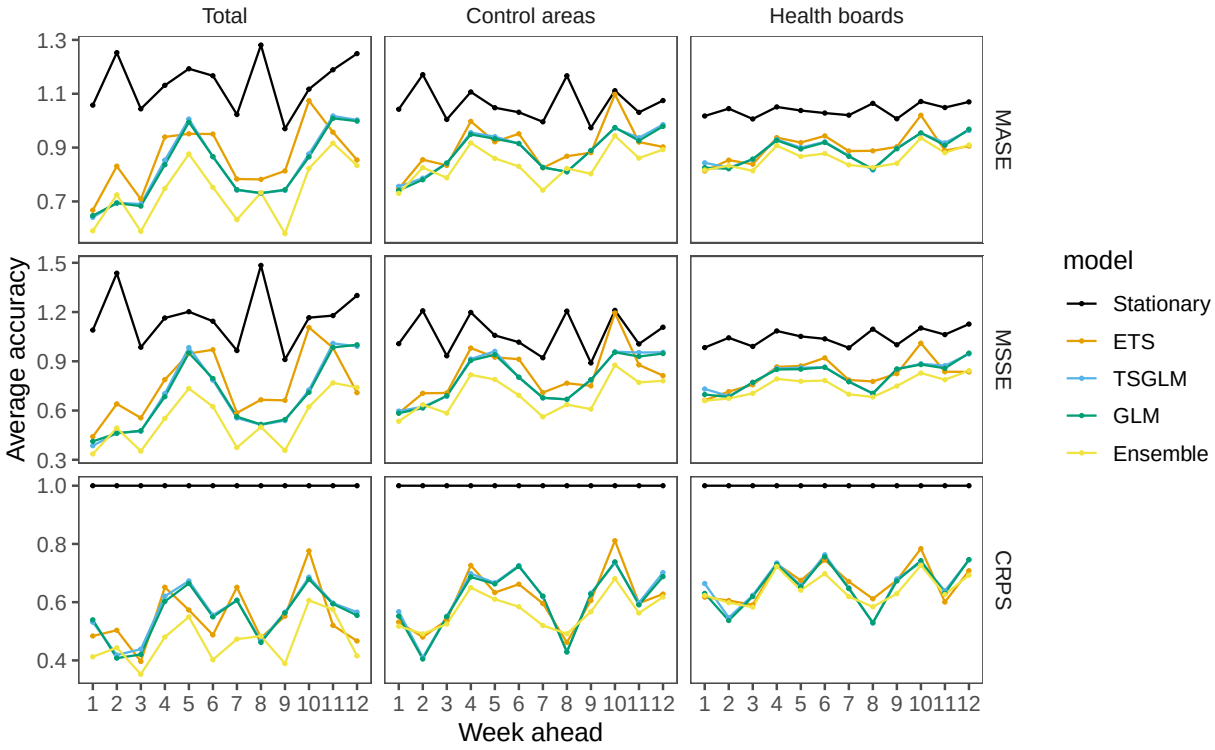


Figure 4: Average accuracy by week for 12 weeks using MinT reconciliation. The total number of days used to calculate the accuracy in the test set is 462. Forecasts are generateled every 42 days, therefore we use 11 samples to calculate the average accuracy. CRPS is relative to a stationary Empirical Cumulative Distribution Function (ECDF). MASE and MSSE are relative to the corresponding values for the training set.

In addition to the overall forecast accuracy presented in Table 3, we also report the point and probabilistic forecast accuracy measures for each forecast horizon in Figure 4. The figure focuses on the hierarchical levels important for decision-making including total, control area, and health board; however, the accuracy could be calculated for any level. We only illustrate the results of the MinT method, given its strong performance described in Table 3. For illustration purposes, we report the average weekly forecast accuracy instead of the daily forecast horizon, as this reduces the visual noise in the figure. Thus, the x-axis shows horizons from week 1 ($h = 1, \ldots, 7$) to week 12 ($h = 78, \ldots, 84$). The forecast horizon from week 7 to week 12 corresponds to the upcoming planning horizon, which is used by planners and decision-makers. For both the point

forecast and distributional accuracy we can see that the ensemble approach performs best across almost all horizons, with the biggest differences at the highest levels of aggregation. It is important to highlight that, all forecasting models outperform the stationary empirical distribution that is used as a benchmark for both point and probabilistic forecasts.

Despite using Poisson regression models to create count distributions of attended incidents for the base forecasts, it is important to note that the reconciled forecast distributions do not maintain a count format. In practical scenarios, there might be a need to use integer forecasts. Count forecast reconciliation is an active area of research, and it would be interesting to explore how our approach could be adapted to generate count-reconciled probabilistic forecasts in future studies. One possible approach to address this is by rounding the forecasts. However, the impact of rounding on forecast accuracy varies depending on the level of hierarchy and the scale of the data. In situations with high-volume demand, the effects of rounding may be negligible, and forecast accuracy calculations can overlook integer effects. On the other hand, in low-volume demand settings, such as forecasts at the bottom level of the hierarchy, integer (rounding) effects may have a more noticeable influence on forecast accuracy.

5.1. An illustration of probabilistic forecast for EMS demand

Figure 5 provides an illustrative example of a probabilistic forecast for future demand, one day ahead, of the total attended incidents in the PO health board. The black dashed line in the graph represents the point forecast (the mean of the distribution), which presents a single numerical estimate for the future number of attended incidents. Due to the complexity of including such plots for the entire hierarchy and 84 days ahead, only one example is presented here for 1 day ahead. However, it is feasible to generate these plots for the entire hierarchy and for any forecast horizon if necessary.

In practice, point forecasts are commonly used, but they have limitations as they ignore the uncertainty associated with the forecast, leading to wrong decisions. In contrast, probabilistic forecasts offer an alternative approach to anticipate future demand. Rather than providing a single value, they assign likelihoods to all possible demand outcomes, acknowledging that different numbers of attended incidents are possible, but with varying likelihoods.

The purpose of probabilistic forecasting, as demonstrated in Figure 5, is to quantify uncertainty. Decisions based on these forecasts could focus on the tails of the distribution: unexpectedly high demand leading to crowding and inefficiency, or unexpectedly low demand resulting in wasted resources. Such forecasts are valuable tools for decision-makers and planners, especially when dealing with low-probability, high-cost situations. Different EMS managements may have varying risk attitudes depending on resource availability, making it crucial to consider the entire distribution when making decisions. For instance, these forecasts enable management to calculate the probability of demand exceeding a certain threshold of available resources (e.g., 90%), which can serve as an informative early warning measure for crowdedness.

It is important to note that while point forecasts and prediction intervals can be obtained from the probabilistic forecasts, the reverse is not possible. A single number cannot be used to directly derive a probabilistic forecast. Prediction intervals, although helpful in indicating possible ranges, do not provide information on the probabilities of low or high demand.

In EMS planning, future demand is just one aspect to consider. Other inputs, such as capacity, should also be treated as probability distributions to adopt a probabilistic approach to planning. To extract valuable insights and make informed decisions from probabilistic forecasts, specialized numerical tools are required, as the forecasts themselves are typically represented as explicit probability density functions or Monte Carlo generators.

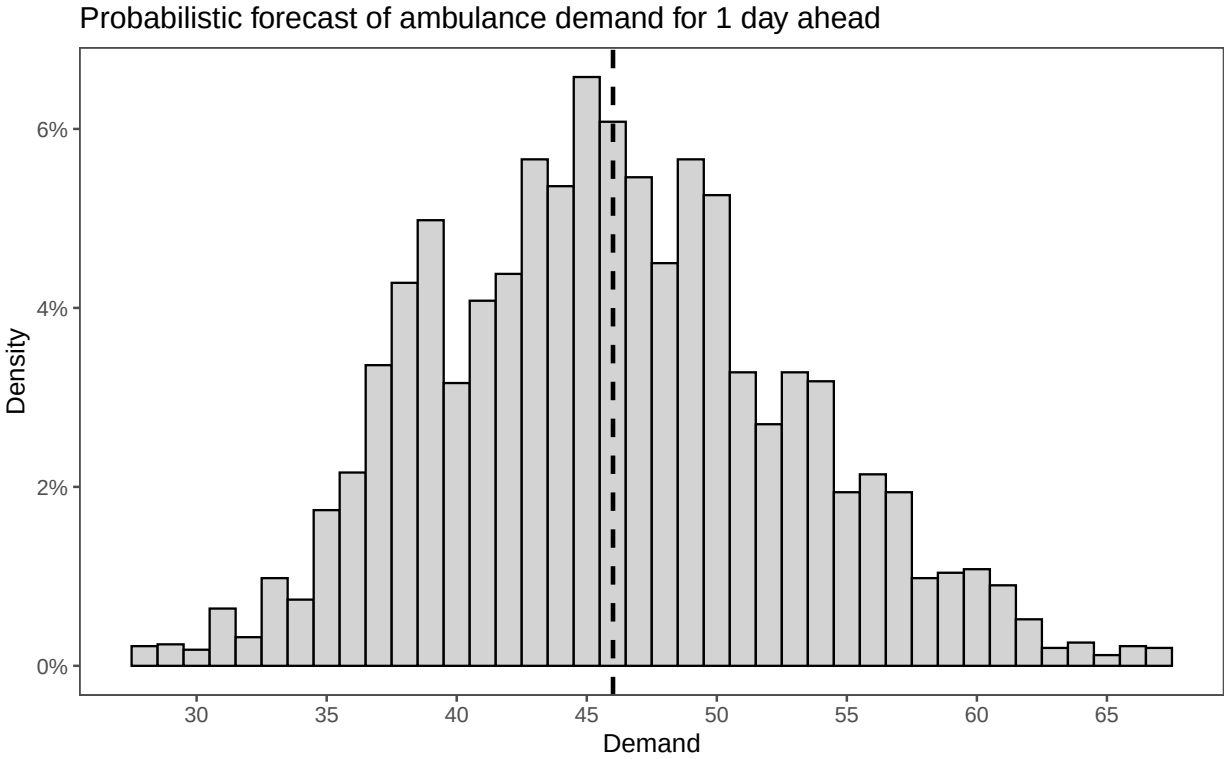


Figure 5: An illustrative example of the forecast distribution of ambulance demand (i.e. total incidence attended) for the PO health board for 1 day ahead. The horizontal axis shows all possible outcomes that may occur, with their likelihood shown on the vertical axis. The dashed line indicates the mean of the distribution.

6. Conclusion

Forecasting problems at Emergency Medical Services often have inherent hierarchical and grouped structures. For example, looking at time series of arrival calls in a clinical desk service, Emergency Department admissions, verified incidents, or attended incidents in a country, they could be disaggregated by various attributes of interest. Total demand in the country could be disaggregated by region, then within each region by health board, within each health board, by station/hospital, and so on down to the postcode area. Alternative structures may arise when attributes of interest are crossed rather than nested. For example, the total demand could be disaggregated by priority (e.g., Red, Amber, Green) or by the nature of incidents. It is also natural to have a mixed structure, for example, the total demand could be disaggregated by priority and by health board.

Despite the inherent hierarchical structure of the forecasting problem in EMS, the common practice is to produce point forecasts for each time series independently. This practice may lead to a lack of coordination and possibly undesirable and conflicting outcomes. Furthermore, due to the asymmetric impact of resource allocation in this area, quantifying forecast uncertainty through probabilistic forecasts is also of value as it enables planners to manage associated risks. In this paper, we investigate the application of hierarchical forecasting methods for producing probabilistic forecasts of daily incidents attended up to 84 days ahead, using different forecasting methods.

Our results indicate that forecast reconciliation in EMS can not only contribute to a more coordinated approach to the planning and decision-making through producing coherent forecasts, but also it can result in forecast accuracy improvements. Our proposed forecasting models, combined with reconciliation

approaches, outperform the empirical distribution benchmark. We show that a substantial forecast improvement can be achieved at higher levels of aggregation by applying forecast reconciliation methods. When a point forecast is of interest at the bottom level of the series, we observe that reconciliation may not improve the forecast accuracy if the bottom series are noisy and lack systematic patterns. However, forecast reconciliation may result in more accurate forecast results for bottom series, if we are interested in the tails of forecast distribution rather than just center measures like mean (i.e. point forecast). Coherent forecasts are also crucial for informing planning activities, and we demonstrate that the proposed models produce coherent forecasts across all forecast horizons. Therefore, we recommend that forecast reconciliation approaches be adopted for routine use in EMS, whenever hierarchical and/or grouped time series data need to be forecasted. Moreover, we found that using an ensemble forecasting model, combining all models developed in this paper, instead of using each individually, works remarkably well for our mixed hierarchical & grouped structure.

Our research establishes a strong basis for future investigations and practical implementation in EMS. Leveraging the hierarchical and grouped structure of demand time series, EMS can use this advanced forecasting framework to generate coherent point and probabilistic forecasts, making the most of all available data at every level of the hierarchy. We acknowledge that a forecast serves a greater purpose beyond its mere existence, ideally enabling the best utility in terms of efficient allocation of medical services, response time, and cost, all informed by the forecast. While we fully appreciate the importance of evaluating forecast quality based on its impact on decision-making processes, it is essential to address the data requirements and methodology involved in measuring this impact. For a comprehensive assessment of the forecasts' implications, access to additional data beyond ambulance demand, covering various decision types, capacity information, constraints in the decision system, and more, becomes necessary. This additional data would offer valuable insights into the specific decisions relying on the forecasts, resulting in a more accurate evaluation of their impact on medical services. Furthermore, measuring the actual impact of forecasts would necessitate an approach that goes beyond forecasting itself. This would involve developing and implementing simulation models capable of replicating decision-making processes based on the forecast inputs. These simulation models would then evaluate the quality of the final decisions, taking into consideration the utilities that are particularly significant in the context of EMS.

Future research can build upon this study in several ways. In future investigations, we aim to explore this avenue by incorporating operational information, simulating decision processes, and assessing the decision impact of this framework on utilities that are significant to the EMS. Linking forecasts with its utilities (e.g. response time, allocation of medical services, resource utilization, cost, etc) can offer an opportunity to maximize benefits through a more holistic planning approach. Aditionally, in our study, we employed Poisson regression models to generate count distributions of attended incidents for the base forecasts. However, it is essential to note that the reconciled forecast distributions are not counts. This observation presents an interesting avenue for future research. Also, the dataset used in this study only includes information on attended incidents, it would be valuable for future research to investigate the impact of failed responses on EMS forecasting, if data on these incidents becomes available. It is also important to note that our methodology for hierarchical time series forecasting can be applied to any time series data in EMS, including those that may include failed responses.

Although our study primarily focuses on Emergency Medical Services, it is essential to emphasize that the framework we propose has broad applicability across various service industries (Ostrom et al., 2010). Our approach is particularly valuable in situations where time series data is structured hierarchically and/or grouped, a common characteristic found in many sectors. This occurs when data can be naturally organized into different levels of hierarchies or when dependencies and relationships exist among entities within the system. For instance, in supply chains (Shugan and Xie, 2000), demand forecasting at different levels of the distribution network, such as regional warehouses or retail stores, is vital for efficient inventory management and minimizing stockouts. Our framework allows the reconciliation of forecasts, ensuring consistency and alignment throughout the supply chain, leading to improved decision-making and operational efficiency. In the financial industry (Kimes and Chase, 1998), where investments span multiple asset classes, geo-

graphical regions, or customer segments, our framework can be applied to forecast portfolio performance, asset allocation, or customer demand. Similarly, in transportation, the framework supports forecasting transportation demand at various levels, optimizing route planning and resource allocation. Likewise, in the hospitality and tourism industry (Dekimpe et al., 2016), it facilitates forecasting demand rates at state, regional, and department levels, enabling strategic pricing, capacity planning, and revenue management for hotels and other travel-related businesses. Additionally, in call centers, accurate call volume forecasting at different levels of the call center hierarchy or grouped structure is crucial for workforce management and resource allocation. Implementing our framework, call centers can generate accurate forecasts for different skill groups, shifts, and locations, ensuring efficient staffing and optimal service levels to meet customer demands.

Reproducibility

R code to produce all results in this paper is available at

References

Al-Azzani, M.A., Davari, S., England, T.J., 2021. An empirical investigation of forecasting methods for ambulance calls-a case study. Health Systems 10, 268–285.

Armstrong, J.S., 2001. Evaluating forecasting methods, in: Armstrong, J.S. (Ed.), Principles of forecasting: a handbook for researchers and practitioners. Kluwer Academic Publishers. chapter 14, pp. 443–472.

Bandara, K., Hyndman, R.J., Bergmeir, C., in press. MSTL: A seasonal-trend decomposition algorithm for time series with multiple seasonal patterns. International J Operational Research .

Boylan, J.E., Goodwin, P., Mohammadipour, M., Syntetos, A.A., 2015. Reproducibility in forecasting research. International Journal of Forecasting 31, 79–90.

Clemen, R.T., 1989. Combining forecasts: A review and annotated bibliography. International Journal of Forecasting 5, 559–583.

Dekimpe, M.G., Peers, Y., van Heerde, H.J., 2016. The impact of the business cycle on service providers: Insights from international tourism. Journal of Service Research 19, 22–38.

Ekström, A., Kurland, L., Farrokhnia, N., Castrén, M., Nordberg, M., 2015. Forecasting emergency department visits using internet data. Annals of Emergency Medicine 65, 436–442.

Faraway, J.J., 2016. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. 2nd ed., Chapman and Hall/CRC.

Feller, W., 1991. An introduction to probability theory and its applications, Volume 2. John Wiley & Sons.

Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annual Review of Statistics and Its Application 1, 125–151.

Grekousis, G., Liu, Y., 2019. Where will the next emergency event occur? Predicting ambulance demand in emergency medical services using artificial intelligence. Computers, Environment and Urban Systems 76, 110–122.

Gul, M., Celik, E., 2020. An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. Health Systems 9, 263–284.

Hermansen, A., Mengshoel, O.J., 2021. Forecasting ambulance demand using machine learning: A case study from Oslo, Norway, in: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 01–10.

Hulshof, P.J., Kortbeek, N., Boucherie, R.J., Hans, E.W., Bakker, P.J., 2012. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. Health Systems 1, 129–175.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., 2022. forecast: Forecasting functions for time series and linear models. R package version 8.19.

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

Hyndman, R.J., Athanasopoulos, G., 2021. Forecasting: principles and practice. 3rd ed., OTexts.

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. Journal of Statistical Software 26, 1–22.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688.

Ibrahim, R., Ye, H., L'Ecuyer, P., Shen, H., 2016. Modeling and forecasting call center arrivals: A literature survey and a case study. International Journal of Forecasting 32, 865–874.

Kang, Y., Hyndman, R.J., Smith-Miles, K., 2017. Visualising forecasting algorithm performance using time series instance spaces. International Journal of Forecasting 33, 345–358.

Kimes, S.E., Chase, R.B., 1998. The strategic levers of yield management. Journal of Service Research 1, 156–166.

Liboschik, T., Fokianos, K., Fried, R., 2017. tscount: An R package for analysis of count time series following generalized linear models. Journal of Statistical Software 82, 1–51.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2022. The M5 competition: Background, organization, and implementation. International Journal of Forecasting 38, 1325–1336.

Ostrom, A.L., Bitner, M.J., Brown, S.W., Burkhard, K.A., Goul, M., Smith-Daniels, V., Demirkan, H., Rabinovich, E., 2010. Moving forward and making a difference: research priorities for the science of service. Journal of Service Research 13, 4–36.
Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., Hyndman, R.J., 2023. Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. European Journal of Operational Research 306, 693–706.
Pennings, C.L., Van Dalen, J., 2017. Integrated hierarchical forecasting. European Journal of Operational Research 263, 412–418.
Rostami-Tabar, B., Ziel, F., 2022. Anticipating special events in emergency department forecasting. International Journal of Forecasting 38, 1197–1213.
Sasaki, S., Comber, A.J., Suzuki, H., Brunsdon, C., 2010. Using genetic algorithms to optimise current and future health planning-the example of ambulance locations. International Journal of Health Geographics 9, 1–10.
Shi, M., Rostami-Tabar, B., Gartner, D., 2022. Forecasting for unplanned care services: A literature review. Working Paper.
Shugan, S.M., Xie, J., 2000. Advance pricing of services and other implications of separating purchase and consumption. Journal of Service Research 2, 227–239.
Stodden, V., Miguez, S., . Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. Journal of Open Research Software 2, p.e21.
Vile, J.L., Gillard, J.W., Harper, P.R., Knight, V.A., 2012. Predicting ambulance demand using singular spectrum analysis. Journal of the Operational Research Society 63, 1556–1565.
Vile, J.L., Gillard, J.W., Harper, P.R., Knight, V.A., 2016. Time-dependent stochastic methods for managing and scheduling Emergency Medical Services. Operations Research for Health Care 8, 42–52.
Villegas, M.A., Pedregal, D.J., 2018. Supply chain decision support systems based on a novel hierarchical forecasting approach. Decision Support Systems 114, 29–36.
Wang, X., Hyndman, R.J., Li, F., Kang, Y., in press. Forecast combinations: an over 50-year review. International Journal of Forecasting .
Wargon, M., Guidet, B., Hoang, T., Hejblum, G., 2009. A systematic review of models for forecasting the number of emergency department visits. Emergency Medicine Journal 26, 395–399.
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, 804–819.
Winkler, R.L., 1972. A decision-theoretic approach to interval estimation. Journal of the American Statistical Association 67, 187–191.
Zhou, Z., 2016. Predicting ambulance demand: Challenges and methods. arXiv preprint arXiv:1606.05363 .
Zhou, Z., Matteson, D.S., 2016. Predicting Melbourne ambulance demand using kernel warping. The Annals of Applied Statistics 10, 1977–1996.