Forecasting short-term hourly Emergency Department arrivals

Bahmna Rostami Tabar*,[a], Jethro Browell**,[b], Ivan Svetunkov**,[c]

[a] *Cardiff Business School, Cardiff University, UK*
[b] *School of Mathematics and Statistics, University of Glasgow, UK*
[c] *Centre for Marketing Analytics and Forecasting, Lancaster University, UK*

Abstract

An accurate forecast of Emergency Department (ED) arrivals by hour of the day is critical to meet patients' demand It will enable planners to match ED staff to number of arrivals, redeploy staff and reconfigure units, if required. This can have many advantages for healthcare staff and the quality of care delivered to patients. In this study, we develop an innovative model based on Generalised Additive Models and an advanced dynamic model based on exponential smoothing, to generate hourly probabilistic forecast of ED arrivals. We compare the forecast accuracy of these models against appropriate benchmarks including TBATS, Poisson Regression, Prophet, and a simple empirical distribution. We use Root Mean Squared Error (RMSE) to examine the point forecast accuracy and the forecast distribution accuracy is assessed using Quantile Bias, PinBall Score and Pinball Skill Score. Our results indicate that the proposed models outperform their benchmarks for both point and probabilistic forecasts. Our developed models can also be generalised to forecast hourly arrivals in other services such as hospitals, ambulance, or clinical desk services.

Key words: Emergency Department, Arrivals, Poisson Regression, Probabilistic Forecasting, Generalised Additive Models, Intermittent Exponential Smoothing

1. Introduction

Forecasting Emergency Department (ED) arrivals is critical for informing staffing and scheduling decisions to meet the needs of patients. An accurate ED demand forecasts contribute to a better decision making process regarding the resources allocation and staffing. This is one of the best ways to optimize resources utilisation and minimise related costs. An accurate forecast of patient arrivals is crucial in ED services to depict various courses of action that can result in massive savings in terms of patient lives. Inability to match the staff with the demand might result in patients overcrowding the system which is a serious problem that causes challenges for the patient flow [10]. Also, it is related with increasing length of stay [24], low patient satisfaction, unexpected return visits to services, increased health care costs, inaccuracy in electronic medical records and other [27].

An accurate forecasting of arrivals by hour of the day enables planners to match staff to meet anticipated patients, reconfigure units and redeploy staff. This has many advantages for both patients, staff and the quality of provided services. Hourly forecasts are required to inform the short-term operational planning for the current and the upcoming shifts of the day. This involves the short-term decision making related to the execution of the delivery process in ED. The combination of an hourly arrival forecast, current staff being occupied, resource availability and waiting times at ED, provide information on the state of the unscheduled care system across the service. Having this full picture enables the delivery managers to focus on the areas that require intervention to enable the most effective delivery of the service to the patients. However, in comparison with lower frequency time series forecasting such as monthly, quarterly and yearly, hourly forecasts are challenging because the noise caused by random variation may overshadow any pattern in the time series. Hourly time

*Corresponding Author
**Equal contribution
Email addresses: rostami-tabarb@cardiff.ac.uk (Bahmna Rostami Tabar), jethro.browell@glasgow.ac.uk (Jethro Browell), i.svetunkov@lancaster.ac.uk (Ivan Svetunkov)

series generally exhibit multiple seasonal cycles of different lengths such as hourly, daily, weekly and yearly. They may also express nonstationarity and their profile may change over time. Therefore, an appropriate forecasting model should take these features into account, to accurately predict hourly demand admissions.

There are few studies that look at forecasting hourly arrivals in ED and other hospital services using historical time series data and/or predictors such as patient characteristics, weather, holidays and public events. These studies use multiple approaches including Exponential Smoothing [31], Autoregressive Integrated Moving Average (ARIMA) [16], Autoregressive Conditional Heteroskedasticity (ARCH) [3], Vector Autoregressive model [21], TBATS [9] and Artificial Neural Networks [16]. However, majority of these studies are limited to only predict future arrivals as a point forecast (a single number), which does not quantify any uncertainty associated with the number of future arrivals. There are few studies that report uncertainty by presenting prediction intervals, no study generates and evaluates the entire forecast distribution of arrivals. Reporting the uncertainty via the forecast distribution is potentially valuable in this setting because the consequences of imperfect staffing are asymmetric. This asymmetry arises because it is preferable to incur a small opportunity cost associated under utilised staff rather than compromise service levels, if staff levels are insufficient. Probabilistic forecasts inform decision-makers about exposure to these risks and potentially enable those risks to be managed more efficiently. Furthermore, if the impact of under and over staffing can be quantified, probabilistic forecasts allow 'optimal' decisions that balance the cost associated with under and over staffing to be calculated. Therefore, in this paper, in addition to generating point forecasts, we also produce and evaluate density forecasts of hourly ED arrivals, comparing several methods for this task. Another drawback of existing studies is that the datasets used are relatively small (e.g. time period of 1-2 year), which make it challenging to report the forecast accuracy using robust approaches such as time series cross validation and such results might not be generalisable. Additionally, most of forecasting methods used in these publications does not consider the full extend of multiple seasonality exhibited by hourly ED arrivals. Moreover, all previous publications referenced in this paper are not fully reproducible as underlying data and code are not available.

In this paper, we aim at filling several gaps, and our contributions to the literature are summarised as follows:

1. We produce probabilistic forecasts, in addition to the point estimation, which quantify uncertainties in future hospital admission, and compare different forecasting methods using a suite of established evaluation metrics;
2. We develop an advanced dynamic model to forecasts ED arrivals based on iETS [33] and ETSX models with a modification for multiple frequencies which produced highly-accurate point forecasts;
3. We develop a novel model to produce probabilistic forecast of ED arrivals based on Generalised Additive Models for Location Scale and Shape, which accounts for i) the bounded and non-Gaussian distribution of arrivals, ii) multiple seasonalities, weather and holiday effects, and iii) variation in forecast uncertainty;
4. We benchmark the accuracy of our model against appropriate models used when multiple seasonality is present, i.e. Prophet, TBATS, Poisson Regression, Exponential Smoothing State Space model (ETS) and a simple empirical distribution of the arrivals;
5. We provide data and code enabling reproduction and refinement of the proposed approach and benchmarks. The proposed approach could also be generalised to forecast hourly requirements in other services such as the number of incidents or call volumes in clinical desk services.

The rest of the paper is organised as following: In section 2, we provide a brief overview of hourly ED arrival forecasting; In Section 3, we present the hourly time series of an ED arrival and use various plots to highlight important patterns. In section 4, we describe the modelling approach and benchmark methods. We then discuss the performance evaluation metrics in section 5; in section 6, we present and discuss our results. Finally, we summarise our findings and present ideas for future research in section 7.

2. Research background: hourly ED forecasting

There are some studies that employ models to forecast admissions and arrivals in Emergency Department. The time granularity considered by these studies spans from hourly to yearly. However, given the focus of the paper, we only discuss the hourly ED forecasting.

Linear regression, ARIMA, and Naïve were used by Hertzum [13] to investigate whether accurate hourly accident and emergency department patient arrivals and occupancy forecasts can be generated using calendar variables. Hertzum [13] study found that patient arrival variation is larger across the hours of the day than across the days of the week and the months of the year. In terms of hour of the day, patient arrivals peaked around noon. For days of the week, Monday is the busiest day while weekends are the quietest ones. July-August are the months with the highest numbers of patient arrivals, while January and February are the months with the lowest numbers. They indicate that regression and ARIMA models performed similarly in modelling patient arrivals, while ARIMA outperformed regression models in modelling accident and emergency department occupancy.

Choudhury and Urena [8] used ARIMA, Holt-Winters, TBATS, and neutral network methods to forecast hourly accident and emergency department arrivals. ARIMA model was selected as the best fit model. Authors claimed that ARIMA has provided high and acceptable hourly ED forecasting accuracy, even outperforming TBATS. Cheng et al. [6] develop an ARIMA model for ED occupancy with a seasonal component and exogenous variables, which is found to outperform a rolling-average benchmark. They also produce prediction intervals, a form of probabilistic forecast, which are found to be well calibrated, a necessary property for such forecasts.

Morzuch and Allen [23] used the Unobserved Components Model (UCM), in which each component of the time series is separately modelled as stochastic. Double-seasonal exponential smoothing and standard Holt-Winters were used to forecast ED arrival for an horizon of 168 hours. The hourly data collected from an ED in Pennsylvania showed no trend, and two seasonal cycles: a within-day and a within-week seasonal cycles. The double seasonal model recorded lower RMSEs for all the 168-hour horizons, which was expected due to the strong hourly seasonality of the time series.

McCarthy et al. [22] employed a Poisson log-linear regression model, including independent variables such as temporal factors (e.g., hour-of-day, day- of-week, type-of-day, season, and calendar year), patient characteristics (i.e., age, gender, insurance status, triage level, mode of arrival, and ambulance diversion status) and climatic factors (i.e., temperature and precipitation) to model patient demand for ED services. The authors produced probabilistic predictions in the form of 50% and 90% prediction intervals for the number of hourly arrivals. Hourly data of ED arrivals in the 1-year study period was modelled and analysed, and it is suggested that the model could be used for forecasting, however, model evaluation was performed in-sample on only one year of data, so unclear how this approach would perform in a forecasting setting or compare to simpler approaches. However, the length of time series is very short (only one year), which does not allow for a rigorous out-of-sample evaluation.

Schweigler et al. [28] conducted an investigation on whether time series methods could accurately generate short-term forecasts of ED bed occupancy. A year-long dataset of hourly ED bed occupancy was collected from three facilities. For each facility, the authors implemented an hourly historical average model, SARIMA model and sinusoidal model with autocorrelated error. The historical average model was based on the mean occupancy for each site, for each hour of the day; while the sinusoidal model was based on 4 parameters: an AR term, a sine coefficient, a cosine coefficient and an intercept. They evaluated the forecast accuracy of four and twelve hours forecast horizon using RMSE and they found that both SARIMA and the sinusoidal models outperformed the historical average (for example, at site 2, the two models improved by 33% the 12-hour forecasts generated by historical average).

Kim et al. [19] compared different univariate and multivariate time series forecasting techniques to forecast patient volume for a Hospital Medicine programme. The study evaluated linear regression, exponential smoothing, ARIMA, SARIMA, Generalized Autoregressive Conditional Heteroskedasticity (GARCH) and Vector Autoregressive (VAR) models to forecast for 4 hours and 24 hours ahead. They used Mean Absolute Percentage Error (MAPE) to report the forecast accuracy. ARIMA model outperformed all the other models.

Table 1 summarize the relevant papers.

Table 1: Summary of studies in forecasting hourly arrivals in Emergency Department

Author	Year	Variable	Horizon	Length	Method	Metric	Probabilistic	Seasonality
Cheng et al.	2021	ED visits	1h to 4 h	1 year	SARIMAX, Holt-Winters, VAR, ARIMA	MSE, MAE, MAPE, Prediction interval coverage	NO	Single
Choudhury and Urena	2020	ED arrivals	1h to 24h	4 years	ARIMA; Holt-winters; TBATS; ANN	RMSE, ME	No	Multiple
Asheim et al.	2019	ED arrivals	3h	5 years	Poisson regression	MAPE	No	Single
Hertzum	2017	ED arrivals	1,2,4,8,24 h	3 years	linear regression; SARIMA; Naïve	MAE, MAPE, MASE	No	Single
Kim et al.	2014	Hospital admission	4h, 24h	3 years	Linear regression; Exponential smoothing; ARIMA; GARCH; VAR	MAPE	No	Single
Cote et al.	2013	ED arrivals	24h	2 years	Fourier regression	R^2, Standard Error	No	Single
Chase et al.	2012	ED CUR	30m 1h, 2h, 4h, 8h, 12h	1 year	Binary regression	NA	No	Single
Schweigler et al.	2009	Bed occupancy	4h, 12h	4 years	Hourly historical average; SARIMA; Sinusoidal model with autocorrelated error	RMSE	No	Single
Jones et al.	2009	ED census	24h	2 years	VAR; Holt winters	MAE	No	Single
McCarthy	2008	ED arrivals	n/a	1 year	Poisson log-linear regression model	Prediction interval coverage	Yes	Multiple
Channouf et al.	2007	Ambulance admission	1h, 3h, 6h, 12h, 13h, 14h, 17h, 23h, 24h	2 years	Regression	RMSE	No	Single
Morzuch and Allen	2006	ED arrivals	168h	3 years	Double Exponential Smoothing; Additive Holt Winter	RMSE	No	Multiple

Asheim et al. [1] developed a Poisson time-series regression model with continuous day-of-week and week-of-year effects to implement a real-time system that could forecast ED arrivals on 1, 2, 3 hours ahead. Measuring the accuracy using the MAPE, Asheim et al. [1] noticed that great improvement happened when time of notification was incorporated into the model, especially in one-hour horizon.

Cheng et al. [7] used one year of ED visits time series to evaluate the performance Rolling Average, SARIMAX, ARIMA, VAR and Holt-Winter to forecast ED occupancy up to 4-hours ahead. The forecast accuracy is evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE) and MAPE for point forecast and coverage for prediction intervals of 80% and 95%. They show that SARIMAX provides more accurately forecast of hourly ED occupancy.

According to the studies mentioned above, it can be said that they have shown complications in forecasting hourly patient accident and emergency department visits and the application of forecasting hourly patients visits is not well established. Some of the studies claimed that the accuracy of forecasting models on hourly accident and emergency department data forecasting model is low compared to the higher forecasting intervals, like daily [5, 13]. While, others mentioned that the accuracy of ED hourly forecast is at the acceptable level [8, 22, 28].

There are few limitations in the literature which encourage us to undertake this research and examine different forecasting approaches. These limitations are summarised as follows:

(i) Current approaches to forecast hourly ED arrivals do not fully consider the feature of data such as multiple seasonal cycles and changing profile over time;
(ii) Almost all research studies produce point forecasts and at best report prediction intervals. There is a lack of studies presenting the entire forecast distribution of hourly ED arrivals that better represent uncertainty of future arrivals, providing a holistic picture of future demand for a planner;
(iii) most studies are not reproducible, as it is almost impossible to reapply the approaches without the help of the authors of those papers;
(iv) studies are limited in terms of the length of historical data used for training purposes and forecast performance evaluation and
(v) some studies in this area lack a rigorous experimental design, i.e. there is no benchmark method or forecast accuracy is not reported.

3. Preliminary analysis

Data used in this study comprises counts of patients' arrival times at one of the largest ED units in the UK between April 2014 and February 2019, extracted from the ED administrative database of the hospital. We aggregated the patients' arrival times to obtain hourly arrivals, which are used in this study. Figure 1 illustrates the distribution of arrivals for each hour of the day and the day of the week. Although the data is noisy, it reveals some systematic patterns.

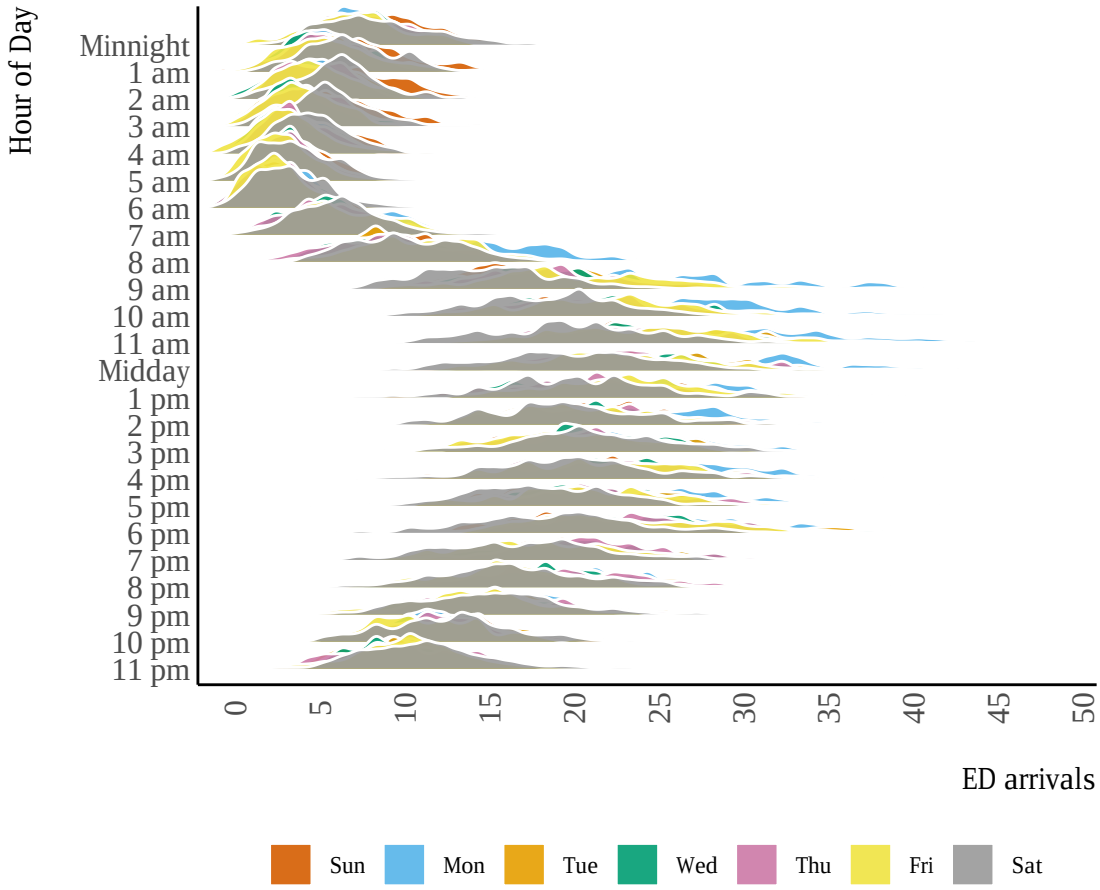

Figure 1: Distrubution of admission per hour and day of the week

It is clear that the number of arrivals has a sub-daily structure, similar to the one summarised by Hertzum [13]. The ED arrivals decrease between mid-night and early morning and then increase until the evening, decreasing after that again. It is also clear that ED service gets systematically more visits on Mondays between 8 a.m. and 5 p.m. Moreover, the number of arrivals around mid-night is slightly higher for Saturday and Sunday.

Figure 1 also highlights that there is significant skewness for almost every hour of the day that varies with time-of-day, and which should be accounted for in forecasting methods. Some skewness might be related to holidays and special events. It is also clear that arrivals are less volatile between mid-night and early morning.

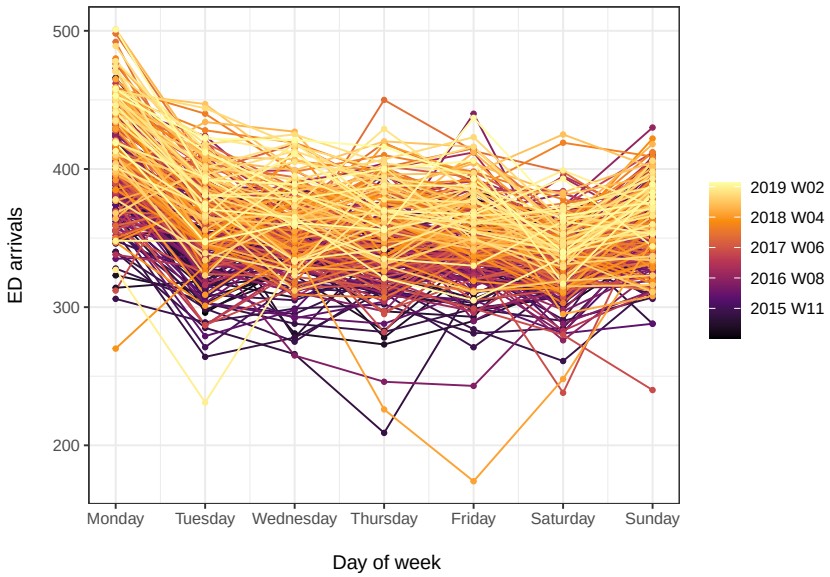

Figure 2: Seasonal plot: day of week arrivals

Figure 3 illustrates the daily seasonal plot, with x-axis representing the day-of-week and y-axis the ED arrivals. There is a line for each week (i.e. 7 days) from Monday to Saturday, and lines are color-coded by week of year to show the weekly cycle. It is clear that there is a large jump in patient arrivals on Monday followed by Saturday each week. This shows that there are significantly more arrivals on Mondays and Saturdays comparign to the rest of the week. This might be due to the closure of General Practitioners outpatient clinics over the weekend.

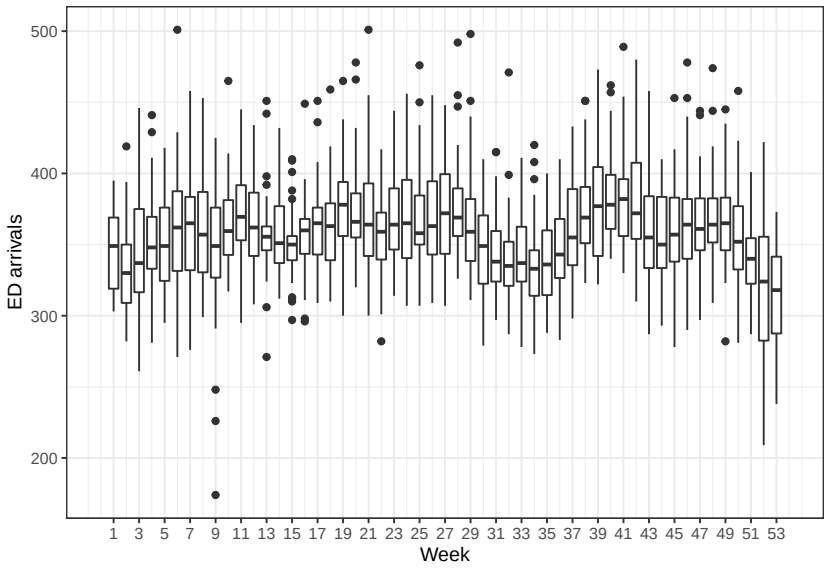

Figure 3: Week of year arrivals

Figure 3 highlights the week of year seasonality in the ED arrivals. We observe that arrivals are significantly lower from week 29 to 35, which corresponds to the summer period. Moreover, the number of arrivals are lower at beginning and the end of the year. The highest arrivals are in week 39-42.

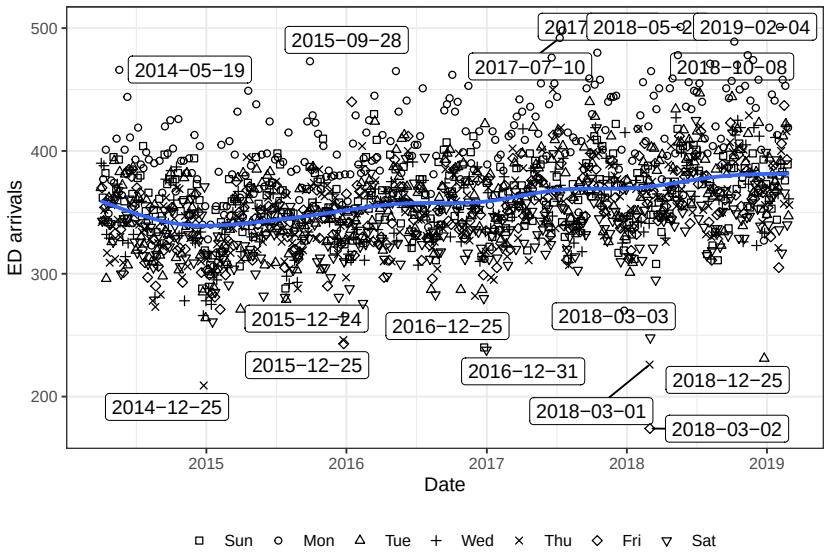

Figure 4: Daily arrivals

Figure 4 presents the time plot of daily arrivals. Each point represents one day, and points are shape-coded by day-of-week to show the weekly cycle. The figure shows more arrivals on Mondays compared to other weekdays. Moreover, we can observe a long-term trend line as well as the significant effect of some holidays. We can see that arrivals near Christmas and New Year's day are significantly lower than other days of the year. Moreover, Figure 4 allows us to identify the impact of special event on arrivals. For instance, we see that the number of arrivals are significantly low for 01-03 of March 2018. These days correspond to the Storm Emma with heavy snowfall that resulted in travel disruption, mass power outages and schools closed in the UK [37].

Based on ED time series analysis illustrated above and literature review, we should consider models that can take the following into account:

- Daily, weekly, and annual seasonalities,
- Long-term trend,
- Calendar events, such as holidays,
- Lags of calendar events to accommodate the potential changes in demand the next day after a holiday,
- Other events, such as sporting fixtures,
- Weather effects, such as temperature and precipitation.

We propose several forecasting models that account for the structures outlined above.

4. Model building

4.1. Naive/Climatology

We start with one of the simplest forecasting approaches used in practice - the one that assumes that in the next few hours, everything will be the same as in the similar hours of a similar day in the past. This is called "Naïve". In our case, given that we need a distribution of values, we will use a modified approach, used in climatology, where the empirical distribution of the hourly arrival time series is used to forecast the future arrival distribution [20].

4.2. Poisson Regression

Regression is one of the most popular forecasting methods that uses explanatory variables to predict a variable of interest (in our case, the ED arrivals). The classical linear regression model is formulated as

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + \epsilon_t, \tag{1}$$

where $\mathbf{x}_t$ is the vector of explanatory variables, $\boldsymbol{\beta}$ is the vector of parameters, ϵ_t is the error term, which is typically assumed to follow Normal distribution with zero mean and a fixed variance, and t is the time index. However, in the context of healthcare and ED arrivals, the assumption of Normality is unrealistic, because the number of admitted patients is integer and non-negative. So the linear regression model should be substituted by some other model. One of the models that is frequently used in practice is the Poisson regression [see for example, 22], which can be summarised as

$$\begin{aligned} y_t &\sim \text{Poisson}\left(\lambda_t\right) \\ \log \lambda_t &= \mathbf{x}_t' \boldsymbol{\beta} \end{aligned} . \tag{2}$$

The logarithm in (2) is needed in order to make sure that the parameter of Poisson distribution is always positive. This model can be estimated via maximisation of the likelihood function based on Poisson mass function . When it comes to selecting explanatory variables for the model, there is no single correct answer, and the decision needs to be done for each specific case. In our experiment, we will only include dummy variables, capturing a variety of calendar events:

1. Hour of day,
2. Day of week,
3. Week of year,
4. Holidays (such as Christmas, New Year etc),
5. 24 hours lags of holidays.

The variables (1)–(3) allow modelling the seasonal patterns on the appropriate level of detail throughout the year, while (4) covers the changes in admittance due to calendar events. Finally (5) is needed in order to capture the potential phenomenon of change in admittance after the holiday (e.g. people might try not to go to hospital on Christmas eve and thus will go the next day). This model assumes that all these effects are deterministic and do not change over time, but the exponentiation in (2) introduces an interaction effect between dummy variables, so that the 3pm on Monday in January will be different from 3pm on Monday in July, although the parameters for hour of day and day of week are fixed and do not change over time. We use `alm()` function from `greybox` package [29] for R [36] for the experiments and denote this model as "Poisson Regression".

4.3. ETS - Exponential Smoothing

Hyndman et al. [17] developed a state space approach for exponential smoothing models, according to which the model can have a set of components, including different types of Error, Trend and Seasonal component (thus *ETS*). Given the popularity of ETS model in forecasting community, we decided to include the basic ETS(A,N,A) model with the seasonal component with frequency 24 (hour of day) as a benchmark. This was done using `adam()` function from `smooth` package [30] for R and denote as *ETS*. This model does not capture the day of week or week of year effects, does not include explanatory variables, but its seasonal component and level change over time. This model is included as a benchmark, only to see how the other models perform in comparison with it.

4.4. Prophet

Prophet is a forecasting procedure created by Facebook [35] that accounts for multiple seasonality, piecewise trend and holiday effects. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. Prophet works well on daily data seen in Facebook. It is robust and automated, making it easy to learn for beginners. The implementation may be less flexible than other methods. The model itself relies on Multiple Source of Error state space model, originally proposed by

Kalman [18]. The model is incorporated using corresponding implementation of the Fable package in R. We use the `prophet()` function from the `fable` package [25]. Note that the input data is assigned with an hourly and daily seasonality.

4.5. TBATS

De Livera et al. [9] proposed a model to deal with time series exhibiting multiple complex seasonalities called "TBATS". It includes a Box-Cox Transformation, ARMA model for residuals and a trigonometric expression of seasonal terms. The latter not only gives the model more flexibility to deal with fractional seasonality but also reduces the parameters of model when the frequencies of seasonalities are high. We fit a TBATS model using the `tbats()` function from the `forecast` package in R [15].

4.6. ADAM: multiple seasonal iETSX

Svetunkov [32] proposed a framework for dynamic models called the Augmented Dynamic Adaptive Model (ADAM). This framework encompasses ARIMA [4], ETS [17] and regression, supporting multiple frequencies, non-normal distributions and intermittent demand [33]. Based on this framework, we use Gamma distribution for ETS(M,N,M) model with frequencies 24 (hour of day) and 168 (hour of week), adding dummy variables for week of year, holidays and lagged holidays. This way we update the hour of day and day of week seasonal indices, keeping the week of year one fixed, thus reducing the number of estimated parameters. Given that the data exhibits randomly occurring zeroes, we use the direct probability model developed by Svetunkov and Boylan [33] to treat those values. This model can be formulated as a set of the following equations:

$$
\begin{aligned}
&y_t = o_t z_t \\
&\log z_t = \log l_{t-1} + \log s_{1,t-24} + \log s_{2,t-168} + \mathbf{x}_t' \boldsymbol{\beta} + \log\left(1 + \epsilon_t\right) \\
&\log l_t = \log l_{t-1} + \log(1 + \alpha \epsilon_t) \\
&\log s_{1,t} = \log s_{1,t-m} + \log(1 + \gamma_1 \epsilon_t) \\
&\log s_{2,t} = \log s_{2,t-m} + \log(1 + \gamma_2 \epsilon_t) \\
&o_t \sim \text{Bernoulli}\left(\mu_{a,t}\right) \\
&a_t = l_{a,t-1}\left(1 + \epsilon_{a,t}\right) \\
&l_{a,t} = l_{a,t-1}(1 + \alpha_a \epsilon_{a,t}) \\
&\mu_{a,t} = \min(l_{a,t-1}, 1)
\end{aligned}
\quad , \tag{3}
$$

where α, β, γ_1, γ_2 and α_a are the smoothing parameters, defining how adaptive the components of the model should be, l_t is the level component for the demand sizes, $s_{1,t}$ and $s_{2,t}$ are the seasonal components, $\boldsymbol{\beta}$ is the vector of parameters for the explanatory variables, o_t is the binary variable, which is equal to one, when demand occurs and to zero otherwise, $l_{a,t-1}$ is the level component for the occurrence part of the model, and $(1 + \epsilon_t) \sim \Gamma(s^{-1}, s)$, where $s = \frac{1}{T}\sum_{t=1}^{T} e_t^2$ is the scale of the distribution. Finally, a_t is an unobservable series, underlying the occurrence part of the model and $(1 + \epsilon_{a,t})$ is an unobservable error term for a_t. Svetunkov and Boylan [33] discuss how to estimate such a model. We expect this model to perform on par with the Poisson regression, potentially outperforming it in some instances, due to the dynamic parts of the model (level and seasonal components). Although the data is integer-valued, we expect that Gamma distribution will be a good approximation for it. If integer-valued quantiles are needed, then rounding up can be done for them (see 8.1 for the explanation). This model is implemented in `adam()` function from `smooth` package [30] for R and is denoted in our experiment as "ADAM-iETSX".

4.7. GAMLSS

If we assume that our predictive distribution follows a given parametric distribution, as in Poisson regression discussed above, the forecasting task becomes ones of predicting the future values of that distribution's parameters. Generalised Additive Models for Location, Scale and Shape (GAMLSS) are distributional regression models where the parameters are modelled as additive functions of explanatory variables. This provides a powerful and flexible framework for probabilistic forecasting, provided that a

suitable distribution and additive model structures can be found. In practice, this means employing expert judgement and experimenting with various distributions and evaluating their suitability using available training data.

Let $F_t(y_t)$ be a predictive cumulative probability distribution of y_t. In a distributional regression context, $F_t(y_t)$ is modelled via a parametric model, $F(y_t|\boldsymbol{\theta}_t)$, where $\boldsymbol{\theta}_t$ is an m-dimensional vector of parameters. In a GAMLSS framework of Rigby and Stasinopoulos [26] the elements $j = 1, ..., m$ of $\boldsymbol{\theta}_t$ are modelled as

$$g_j(\theta_{j,t}) = \mathbf{A}_{j,t}\boldsymbol{\beta}_j + \sum_i f_{j,i}(\boldsymbol{x}_t^{S_{j,i}}), \quad \text{for} \quad j = 1, \ldots, m, \tag{4}$$

where g_j is a monotonic link function, $\mathbf{A}_{j,t}$ is the t-th row of the design matrix $\mathbf{A}_j$, $\boldsymbol{\beta}_j$ is a vector of regression coefficients, $\boldsymbol{x}_t$ is a d-dimensional vector of covariates and $S_{j,i} \subset \{1, \ldots, d\}$ **is ....** If $S_{j,i} = \{1, 3\}$, then following our notation $\boldsymbol{x}_t^{S_{j,i}}$ is a two dimensional vector formed by the first and third elements of $\boldsymbol{x}_t$. Each $f_{j,i}$ is a smooth function, constructed as

$$f_{j,i}(\boldsymbol{x}^{S_{j,i}}) = \sum_{k=1}^{K_{j,i}} b_k^{ji}(\boldsymbol{x}^{S_{j,i}})\beta_k^{ji}, \tag{5}$$

where b_k^{ji} are spline basis functions of dimension $|S_{j,i}|$, while β_k^{ji} are regression coefficients. The smoothness of each $f_{j,i}$ is controlled via ridge penalties, the definition of smoothness being dependent on the type of effect and penalty being used. See Wood [38] for a detailed introduction to *GAM/GAMLSS* models, smoothing splines bases and penalties.

As our data are counts, the natural starting point is the Poisson distribution with an additive model for $\log \lambda_t$ of the form

$$\log(\lambda_t) = \sum_{i=1}^{7} \beta_i \delta(D_i(t) - i) + \sum_{j=1}^{7} D_j(t) f_j(H(t)) + t f_{\mathrm{Y}}(Y(t)) + f_{\mathrm{Temp}}(Y(t), C_t) \quad . \tag{6}$$

The functions $H(t)$, $D(t)$ and $Y(t)$ return the hour of the day (1–24), day of the week (1–7), and day of the year (1-366) at time t, respectively, and C_t is the temperature at time t.

However, experiments on the training data reveal that calibration of forecasts based on the Poisson distribution is poor, suggesting that the shape of the distribution is unsuitable for the present application. In particular, we observe that forecast uncertainty appears to vary depending on the time of day and possibly other explanatory variables. Therefore, we consider more flexible, two parameter distributions so that we are able to specify additive models for both location and scale parameters, specifically the truncated Normal distribution, with truncation at 0. The resulting density forecasts are given by

$$F_t(y_t, \mu_t, \sigma_t) = \frac{\Phi\left(\frac{y_t - \mu_t}{\sigma_t}\right) - \Phi\left(\frac{-y_t}{\sigma_t}\right)}{1 - \Phi\left(\frac{-y_t}{\sigma_t}\right)} \tag{7}$$

with additive models

$$\mu_t = \sum_{i=1}^{1} 0\beta_i D_i^+(t) + \sum_{j=1}^{1} 0 D_j^+(t) f_j(H(t)) + t + f_{\mathrm{Y}}(Y(t)) + f_{\mathrm{Temp}}(Y(t), C_t)$$

$$log(\sigma_t) = \sum_{i=1}^{1} 0 D_i^+(t) f(H(t))$$

for the mean and variance parameters. We also performed experiments with the truncated t distribution and negative binomial distribution, but these did not results in forecasts as well calibrated as the truncated normal.

5. Forecast performance evaluation

In order to assess performance of models, we evaluate predictive quantiles at probability levels 0.05 to 0.95 in steps if 0.05, and conditional expectations for 0 to 48 hours ahead produced by each model. We forecast upto 48 hours because this is the operational horizon in the ED, for which it is possible to make short-term changes in the shifts for nurses and doctors. The forecasts are produced every 12 hours for the holdout of 365 days in a rolling origin fashion [34], resulting in 727 origins. Based on these values, several error measures are calculated to evaluate the performance of models in terms of specific quantiles and in terms of expectation. The latter is measured via Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{h}\sum_{j=1}^{h} e_{t+j}^2}, \tag{8}$$

where h is the forecast horizon and e_{t+j} is the point forecast error j steps ahead.

The objective of density forecasts is to be as sharp as possible while remaining reliable/calibrated [12]. A forecast is said to be sharp if the predictive distribution has a relatively small spread, indicating low uncertainty, which is valuable to decision makers provided the forecast is calibrated. Calibration, also called reliability, is the property that forecast probabilities match the observed frequency of realisations. If a forecast is calibrated, then, for example, 20% of observations should fall below the $\alpha = 0.2$ predictive quantile (with some tolerance based on the finite sample size). This property is necessary for forecast probabilities to be used in quantitative decision-making. Calibration is typically evaluated visually using reliability diagrams, which plot the nominal coverage, α, against observed frequency mean ($\mathbf{1}(y_t \leq q_{\alpha,t})$). We use several scores to assess the quantile performance of models.

First, in order to measure quantile performance, we need to calculate the pinball score, which is a strictly proper score used to evaluate quantile forecasts and is the discrete form of the Continuous Rank Probability Score [16]. It rewards sharpness and penalises mis-calibration, so measures all-round performance, however, calibration should still be verified separately. Furthermore, The Pinball Score for an individual quantile matches the loss function minimised in quantile regression model. The Pinball Score is given by

$$\text{Pinball} = \frac{1}{T|\mathcal{A}|}\sum_{\alpha\in\mathcal{A}}\sum_{t=1}^{T}\left(q_{\alpha,t} - y_t\right)\left(\mathbf{1}(y_t \leq q_{\alpha,t}) - \alpha\right), \tag{9}$$

where $\mathcal{A} = \{0.05, 0.1, ..., 0.95\}$ is the set of quantiles being estimated.

To compare model performance, and the significance of any apparent difference in performance, we will use skill scores, which can be calculated for any metric via:

$$\text{Skill} = \frac{M_{\text{ref}} - M}{M_{\text{ref}}} \tag{10}$$

where M is the metric's value for the method being considered, M_{ref} is the metric's value for a reference method. The skill score show us by how many percent the reference approach is worse than the one under consideration. We will use bootstrap re-sampling of skill scores to determine if apparent differences in forecast performance (i.e. positive or negative skill) are significantly different from zero [11]. Here we use the best performing simple benchmark, Climatology (explained in Subsection 4.1), as the reference model, and employ a block-bootstrap with blocks of length 24h in order to account for temporal correlations of the underlying data [14, 2].

Finally, we have calculated the computational time for one iteration on the first rolling origin to compare the speed of each function. All functions were re-estimated on each iteration. ADAM and Poisson regression estimated the parameters taking as the pre-initials the ones obtained in the initial model application to the data in the first origin. This allowed to speed up the computation for these two models. The initial estimation of ADAM took approximately one hour and 25 minutes. Each step in the experiment took the time shown below.

6. Results

The data is portioned into training (from 2014-04-01 to 2018-02-28) and test (from 2018-03-01 to 2019-02-28) sets, with all model development and hyper-parameter tuning performed using training data only. The rolling origin advances in 12 hour steps, and the forecast horizon is set for 48 steps ahead.

Figure 5 presents pinball score aggregated across forecasting horizons for each quantile. It shows that the difference in performance among the models mainly comes from the middle of the distribution, and somewhat form the upper tail. There is very little difference in performance for the lower tail. This is kind of interesting, and reassuring that the better models are batter at probabilities that matter more to decision makers.

Probabilistic forecasts are evaluated following the principle of *sharpness subject to calibration*, meaning that the sharper forecast is prefered provided that it is calibrated. Mis-calibrated forecasts are unsuitable for use in decision-making so should be excluded. Calibration is evaluated visually in Figure 6, which highlights a systematic negative bias accross all probability levels in many models, with only the truncated normal and t family GAMLSS models (NOtr-1, NOtr-2, Ttr-2) and ADAM-iETSX models showing good calibration across most probability levels. Notably, both benchmarks exhibit negative quantile bias as they struggle to capture the long term trend of increasing attendance. For a user, this could result in poor staffing decisions as historic data fail to accurately charecterise present conditions.

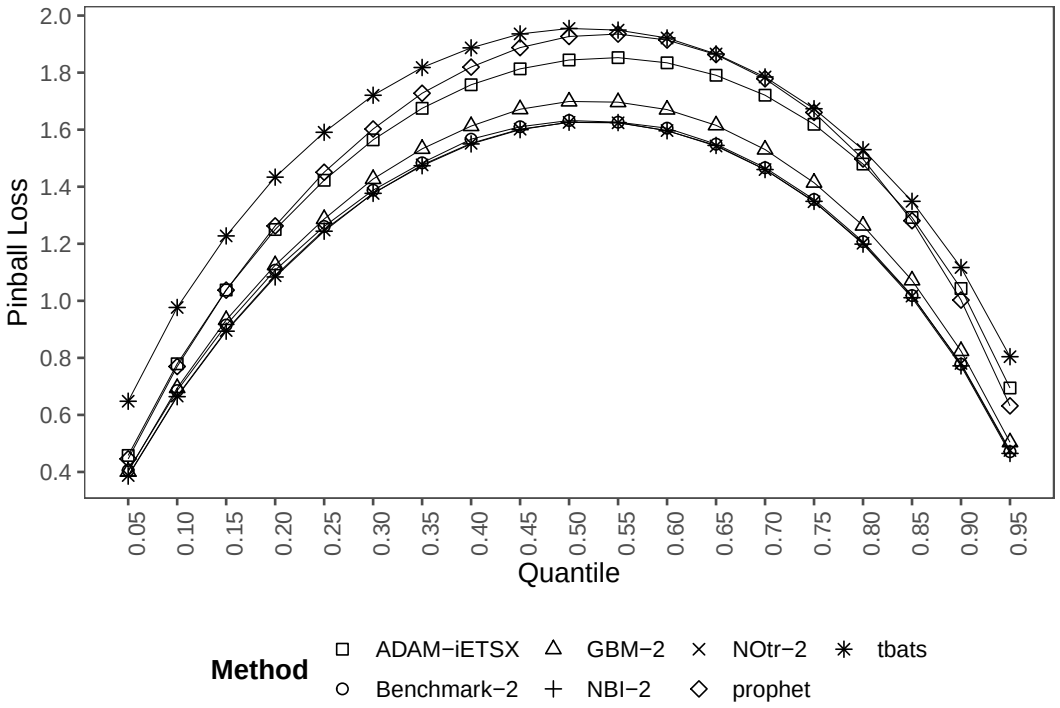

Figure 5: Pinball loss values over different quantiles.

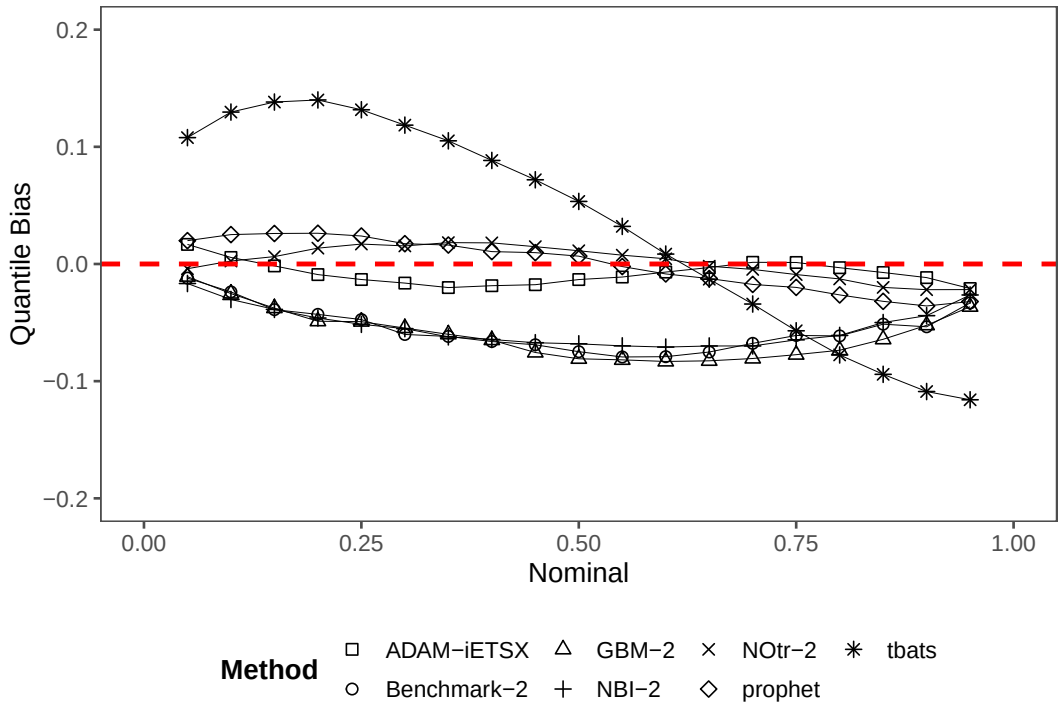

Figure 6: Quantile bias vs the nominal quantiles.

Table 2: Summary of studies in hourly emergency care forecasting

Method	Quantile Bias	Pinball	RMSE	Time
ADAM-iETSX	0.0104673	1.417260	0.0896228	92.9348605
ADAM-iETSX-Ceiling	0.0104673	1.419490	NA	NA
Benchmark-1	0.1047874	1.254491	1.0042634	0.3874450
Benchmark-2	0.0557392	1.217429	0.2592800	0.0947247
ETS	0.0194389	1.434862	0.0121247	10.7175205
GBM-2	0.0600153	1.261690	1.7770897	602.4317496
LinearRegression	NA	NA	NA	3.1543500
NBI-2	0.0540725	1.206241	0.3830272	NA
NOtr-1	0.0098967	1.222583	0.2675957	451.6620471
NOtr-2	0.0118522	1.208561	0.2675957	86.5895462
Poisson-1	0.0372137	1.204920	0.0095263	1.4763353
Poisson-2	0.0373884	1.188109	0.0082932	5.0768588
Poisson-2-I	0.0523937	1.222114	0.3770081	NA
Regression-Poisson	0.0929416	1.293524	0.8490258	67.1401641
Ttr-2	0.0140221	1.210108	0.3324146	956.5532849
fasster	0.1862190	35.051012	127.7504023	NA
prophet	0.0193799	1.447037	0.2955460	20.6755021
qreg-1	0.0643451	1.340557	NA	NA
tbats	0.0855702	1.536080	0.4859770	273.0558176

Evaluation metrics from the test period are presented in Table 2. They are ordered by Quantile Bias. The five models identified above have a Quantile Bias of 0.014 or less, which is substantially lower than the next group of forecast with Quantile Biases of 0.037 and above, ETS being the only

exception with a value of 0.019.

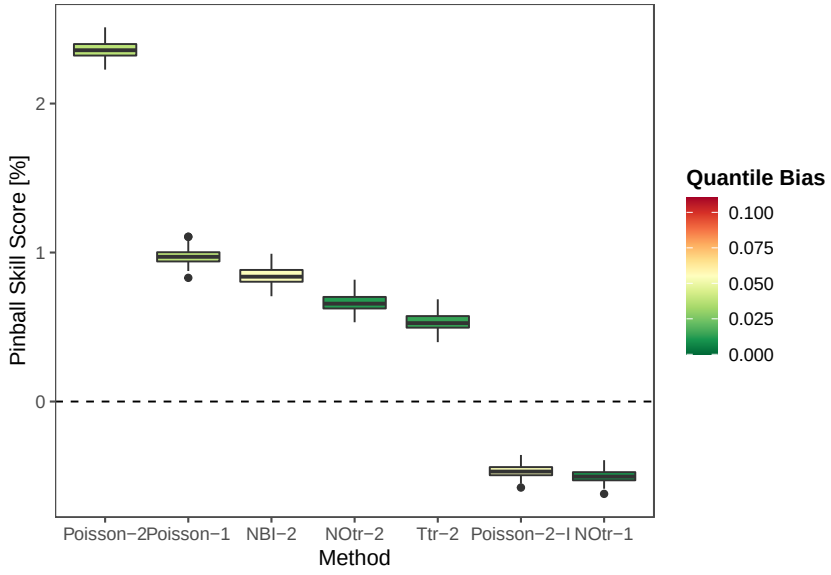

Figure 7: Skill score ...

Figure 7 illustrates the trade-off between calibration and pinball skill.

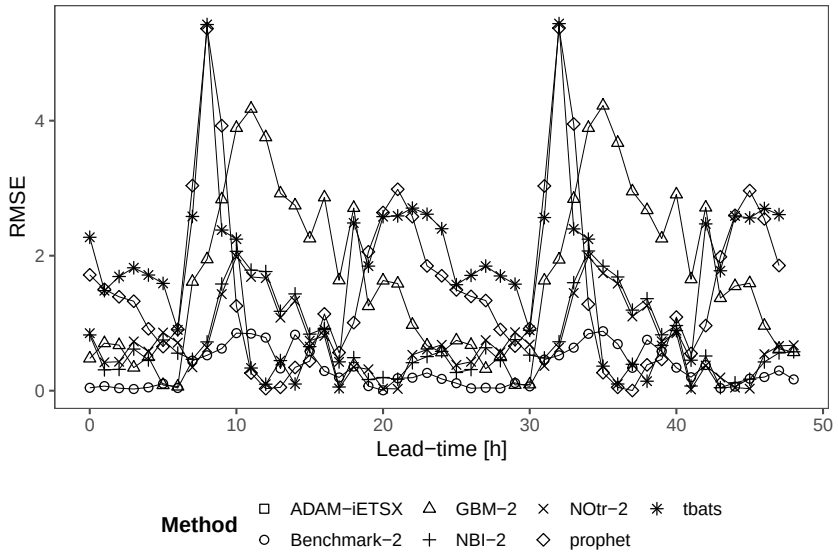

Figure 8: RMSE ...

Figure 8 reports the RMSE for each forecast horizon. It illustrates the times of day that are harder to predict – morning pick-up and afternoon peak.

One more thing to notice is that the ADAM-iETSX model with rounded up quantiles did not perform better than the simpler one with continuous ones. This implies that the rounding is not necessary in general, but if integer values are needed (for example, to decide how many nurses to have), then using the continuous model and then rounding up the quantiles could be considered as a reasonable strategy.

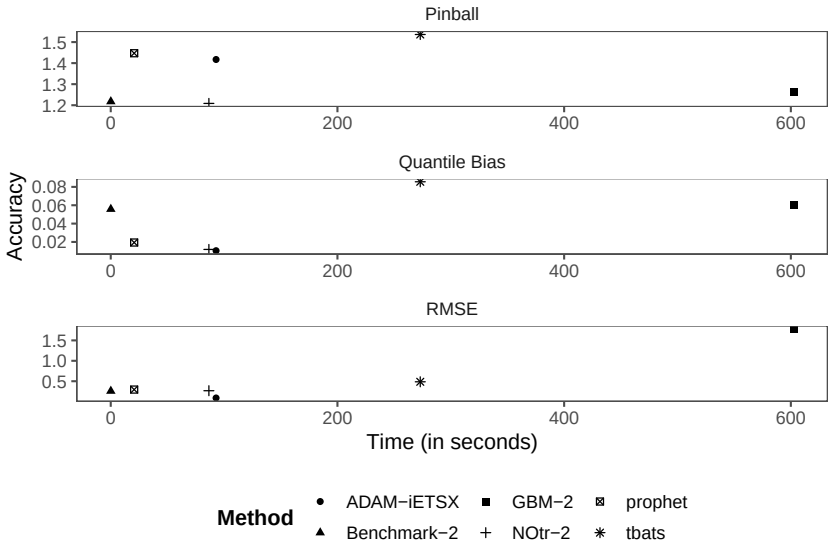

Figure 9: Running time vs. forecast accuracy

Figure 9 reports the forecast accuracy of each approach versus the computational time required to generate the forecast for a given forecasting horizon of 48 hours.

7. Conclusion

Short-term forecasting of arrivals at emergency departments is an important element of staff and resource management in hospitals. Furthermore, due to the asymmetric impact of having an excess of shortage of resources, especially in emergency departments, quantifying forecast uncertainty is also of value as it enables planners to manage associated risks. Here, we have developed methods for producing probabilistic forecast of hourly arrivals up to 48 hours ahead, comparing different state-of-the-art approaches.

Two approaches produced highly accurate, calibrated probabilistic forecasts, one time series method and one based on distributional regression. The first is ADAM-iESTX, which is an extension of exponential smoothing incorporating seasonality and assuming a Gamma predictive distribution. The second, labeled NOtr-2, regressed the two parameters of a truncated (at zero) normal distribution on features date and time features, and temperature. Both approaches produced calibrated probabilistic forecasts, but the point prediction produced by ADAM-iESTX produced forecasts had a lower RMSE than NOtr-2, and NOtr-2 produced forecasts with a lower pinball score. This suggests that the latter may be preferred if the whole distribution is used in decision-making.

Having compared the performance of a wide range of methods, we make the following observations: the choice of distribution assumed for probabilistic forecasts, and choice of model features, are as if not more important than the type of model employed; methods based on quantile regression, which do not assume a parametric distribution for forecasts, do not perform as well as those based on parametric distributions; and the best performing models handled the non-negative and skewed nature of the data automatically without need for post-processing. These observations reflect the characteristics of the data, which is representative of ED arrivals, but determining the extent to which they generalise is beyond the scope of this article. Furthermore, methods based on continuous valued distributions are not adversely affected by the fact that the data are integer-valued. Rounding up predictive quantiles to the next integer does not make predictions worse.

Finally, we have found that out-of-the-box models, those which require minimal tuning or manual development, do not perform as well as well-considered statistical methods. The popular TBATS, Prophet and Gradient Boosting Machine algorithms perform poorly compared to ADAM-iETSX and NOtr-2, and even the benchmarks. Of the models requiring a modest amount of user input and expertise, exponential smoothing (ETS) was found to perform well. ETS produces reasonably well

calibrated forecasts, in contrast to the benchmarks which were poorly calibrated, and highly accurate point forecasts. However, its probabilistic forecasts were considerably worse than NOtr-2 in terms of pinball score.

Probabilistic forecasting opens the door to more sophisticated resource management in healthcare settings by providing decision-makers with uncertainty information and enabling quantitative risk management. Linking forecasts of arrivals with upstream (ambulance call-outs) and downstream (length of stay, medical outcomes) analytics offers an opportunity to improve forecast skill, and may also be necessary to maximise benefits though more holistic decision-making.

Further research is required to investigate the practical benefits of the probabilistic forecasts in the healthcare and how they should be used to inform planning and decision making. This may require employing discrete simulation or new-vendor problem. While this study has focused on the hourly short-term forecasting, producing longer term daily forecast (e.g. 180-270 days ahead) is often required by planners, to support winter planning in ED and Ambulance services which requires more investigation. Moreover, more research is required to forecast other important variables such as length of stay, bed occupancy and waiting time, in addition to patient arrivals and admissions. This may require considering the dynamics among various services including General Practitioners, Emergency Departments, Ambulance and Fire & Rescue services.

8. Appendices

8.1. Quantiles of rounded up random variables

Before proceeding with the proof we need to give the definition of the quantiles of the continuous and rounded up random variables:

$$P\left(y_t < k\right) = 1 - \alpha, \tag{11}$$

and

$$P\left(\lceil y_t \rceil \leq n\right) \geq 1 - \alpha, \tag{12}$$

where n is the quantile of the distribution of rounded up values (the smallest integer number that satisfies the inequality (12)) and k is the quantile of the continuous distribution of the variable.

In order to prove that $n = \lceil k \rceil$, we need to use the following basic property:

$$\lceil y_t \rceil \leq n \iff y_t \leq n, \tag{13}$$

which means that the rounded up value will always be less than or equal to n if and only if the original value is less than or equal to n. Taking into account (13), the probability (12) can be rewritten as:

$$P\left(y_t \leq n\right) \geq 1 - \alpha. \tag{14}$$

Note also that the following is true:

$$P\left(\lceil y_t \rceil \leq n - 1\right) = P\left(y_t \leq n - 1\right) < 1 - \alpha. \tag{15}$$

Taking the inequalities (11), (12), (14) and (15) into account, the following can be summarised:

$$P\left(y_t \leq n - 1\right) < P\left(y_t < k\right) \leq P\left(y_t \leq n\right), \tag{16}$$

which is possible only when $k \in (n-1, n]$, which means that $\lceil k \rceil = n$. So the rounded up quantile of continuous random variable y_t will always be equal to the quantile of the descritised value of y_t:

$$\lceil Q_\alpha(y_t) \rceil = Q_\alpha\left(\lceil y_t \rceil\right). \tag{17}$$

It is also worth noting that the same results can be obtained with the floor function instead of ceiling, following the same logic. So the following equation will hold for all y_t as well:

$$\lfloor Q_\alpha(y_t) \rfloor = Q_\alpha\left(\lfloor y_t \rfloor\right). \tag{18}$$

References

[1] Andreas Asheim, Lars P Bache-Wiig Bjørnsen, Lars E Næss-Pleym, Oddvar Uleberg, Jostein Dale, and Sara M Nilsen. Real-time forecasting of emergency department arrivals using prehospital data. *BMC emergency medicine*, 19(1):42, 2019.

[2] C. Bergmeir, R.J. Hyndman, and J.M. Benítez. Bagging exponential smoothing methods using stl decomposition and box-cox transformation. *International Journal of Forecasting*, 2016.

[3] Tim Bollerslev, Robert F Engle, and Daniel B Nelson. Arch models. *Handbook of econometrics*, 4: 2959–3038, 1994.

[4] George Box and Gwilym Jenkins. *Time series analysis: forecasting and control*. Holden-day, Oakland, California, 1976.

[5] Justin Boyle, Melanie Jessup, Julia Crilly, David Green, James Lind, Marianne Wallis, Peter Miller, and Gerard Fitzgerald. Predicting emergency department admissions. *Emergency Medicine Journal*, 29(5):358–365, 2012.

[6] Qian Cheng, Nilay Tanik Argon, Christopher Scott Evans, Yufeng Liu, Timothy F. Platts-Mills, and Serhan Ziya. Forecasting emergency department hourly occupancy using time series analysis. *The American Journal of Emergency Medicine*, 48:177–182, 2021. ISSN 0735-6757. doi: https://doi.org/10.1016/j.ajem.2021.04.075. URL https://www.sciencedirect.com/science/article/pii/S0735675721003600.

[7] Qian Cheng, Nilay Tanik Argon, Christopher Scott Evans, Yufeng Liu, Timothy F Platts-Mills, and Serhan Ziya. Forecasting emergency department hourly occupancy using time series analysis. *The American Journal of Emergency Medicine*, 48:177–182, 2021.

[8] Avishek Choudhury and Estefania Urena. Forecasting hourly emergency department arrival using time series analysis. *British Journal of Healthcare Management*, 26(1):34–43, 2020.

[9] Alysha M De Livera, Rob J Hyndman, and Ralph D Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527, 2011.

[10] Robert W Derlet. Overcrowding in emergency departments: increased demand and decreased capacity. *Annals of emergency medicine*, 39(4):430–432, 2002.

[11] Bradley Efron. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, 1981. ISSN 00063444.

[12] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2): 243–268, 2007.

[13] Morten Hertzum. Forecasting hourly patient visits in the emergency department to counteract crowding. *The Ergonomics Open Journal*, 10(1), 2017.

[14] GS Hongyi Li and Maddala. Bootstrapping time series models. *Econometric reviews*, 15(2):115–158, 1996.

[15] Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *forecast: Forecasting functions for time series and linear models*, 2020. URL http://pkg.robjhyndman.com/forecast. R package version 8.12.

[16] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2021.

[17] Rob J. Hyndman, Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder. *Forecasting with Exponential Smoothing*. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-71916-8.

[18] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35, 1960. ISSN 00219223. doi: 10.1115/1.3662552.

[19] Kibaek Kim, Changhyeok Lee, Kevin O'Leary, Shannon Rosenauer, and Sanjay Mehrotra. Predicting patient volumes in hospital medicine: A comparative study of different time series forecasting methods. *Northwestern University, Illinois, USA, Scientific Report*, 2014.

[20] Josselin Le Gal La Salle, Mathieu David, and Philippe Lauret. A new climatology reference model to benchmark probabilistic solar forecasts. *Solar Energy*, 223:398–414, 2021.

[21] Helmut Lütkepohl. Vector autoregressive models. In *Handbook of research methods and applications in empirical macroeconomics*. Edward Elgar Publishing, 2013.

[22] Melissa L McCarthy, Scott L Zeger, Ru Ding, Dominik Aronsky, Nathan R Hoot, and Gabor D Kelen. The challenge of predicting demand for emergency department services. *Academic Emergency Medicine*, 15(4):337–346, 2008.

[23] Bernard J Morzuch and P Geoffrey Allen. Forecasting hospital emergency department arrivals. 26th Annual Symposium on Forecasting, Santander, Spain., 2006.

[24] GUL Muhammet and Ali Fuat Guneri. Forecasting patient length of stay in an emergency department by artificial neural networks. *Journal of aeronautics and space technologies (Havacilik ve uzay teknolojileri dergisi)*, 8(2):1–6, 2015.

[25] Mitchell O'Hara-Wild, Rob Hyndman, Earo Wang, and Gabriel Caceres. *fable: Forecasting models for tidy time serie*, 2020. URL https://fable.tidyverts.org/. R package version 0.2.1.

[26] R A Rigby and D M Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):554–507, 2005.

[27] Bahman Rostami-Tabar and Florian Ziel. Anticipating special events in emergency department forecasting. *International Journal of Forecasting*, 2020.

[28] Lisa M Schweigler, Jeffrey S Desmond, Melissa L McCarthy, Kyle J Bukowski, Edward L Ionides, and John G Younger. Forecasting models of emergency department crowding. *Academic Emergency Medicine*, 16(4):301–308, 2009.

[29] Ivan Svetunkov. *greybox: Toolbox for Model Building and Forecasting*, 2021. URL https://github.com/config-i1/greybox. R package version 1.0.2.

[30] Ivan Svetunkov. *smooth: Forecasting Using State Space Models*, 2021. URL https://github.com/config-i1/smooth. R package version 3.1.1.

[31] Ivan Svetunkov. Forecasting and analytics with adam. OpenForecast, 2021. URL https://openforecast.org/adam/. (version: [current date]).

[32] Ivan Svetunkov. Forecasting and analytics with adam. OpenForecast, 2021. URL https://openforecast.org/adam/. (version: 22.06.2021).

[33] Ivan Svetunkov and John E. Boylan. Multiplicative state-space models for intermittent time series. 2019.

[34] Leonard J. Tashman. Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000. ISSN 01692070. doi: 10.1016/S0169-2070(00)00065-0.

[35] Sean J Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

[36] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

[37] BBC Wales. Wales returning to normal as snow thaws and temperatures rise. https://www.bbc.co.uk/news/uk-wales-43274844. Accessed: 2022-01-11.

[38] Simon N Wood. *Generalized additive models: an introduction with R.* chapman and hall/CRC, 2017.