Forecasting short-term hourly Emergency Departement arrivals

Author1[*,a], Jethro Browell[**,1,1,1], Ivan Svetunkov[**,1,1,1]

[a] *Cardiff business school, 3 Colum Drive, CF10 3EU, Cardiff*
[b] *adress2*
[c] *adress3*

Abstract

The Objective of this work would be to propose a new methodology to forecast short-term hourly forecasting for urgent and emergency care.

1. Introduction

Forecasting Emergency Department (ED) arrivals is a critical input to inform staffing and scheduling decisions to meet the needs of patients. An accurate ED demand forecast contributes to a better decision making process regarding the resources needed to provide services to the number and type of patients requiring hospital services. This is one of the best ways to optimize staffing utilization and related costs. Most current practice to optimise personnel scheduling follows the general approach originally presented in Vile et al. [32], which recommends that the following steps be taken to roster employees: (i) forecast demand; (ii) convert demand forecasts into staffing requirements; (iii) schedule shifts optimally; and (iv) assign employees to shifts. An accurate demand forecasting 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 results in patient overcrowding in the system which is a serious problem that causes challenging situations on patient flow [8]. Also, it is related with increasing length of stay [18], low patient satisfaction, number of patients left the hospital without seen by staff, unexpected return visits to services, increasing health care costs, inaccuracy in electronic medical record, and reported waiting times without incurring last-minute expenses, such as overtime or supplemental staffing [22].

There exists a large literature on forecasting ED arrivals that use various methods to forecast annual[28] monthly,[4, 15]. daily[22, 20] and subdaily [23, 5] arrivals. In this paper, we only focus on forecasting hourly arrivals. In comparison to 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.

An accurate demand forecast by hour of the day enables planners to match staff to meet anticipated patients, reconfigure units and redeploy staff. This will have 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 forecast demand, 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. Moreover, there are often unplanned events, such as walk-in patients, extended consultation times, during a day or a week. Staff capacities may be adjusted based on the predicted demand fluctuations by using part-time, on-call nurses, staff overtime and voluntary absenteeism. That may also require to adjust the scheduling.

[*] Corresponding Author
[**] Equal contribution
Email addresses: email1@example.com (Author1), jethro.browell@glasgow.ac.uk (Jethro Browell), i.svetunkov@lancaster.ac.uk (Ivan Svetunkov)

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 , climate factors , holidays. These studies use multiple approaches including Exponential Smoothing [REF], ARIMA families [], ARCH [REF], Vector Autoregressive [REF], TBATS [] and ANN[]. Hourly time 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. This is currently missing in the literature and we fill this gap by examining various forecasting models capable of considering these features. All these studies only generate the estimated future arrivals as a point forecast (a single number), which does not include any uncertainty around the number of arrivals. However, reporting uncertainty is critical because the consequences of imperfect staffing are asymmetric. Therefore, probabilistic forecasts are necessary to make decisions that balance the cost associated with under and over staffing. This asymmetry arises because it is preferable to incur a small opportunity cost associated under utilised staff rather than lower service levels if staff levels are insufficient. The lack of probabilistic forecast is one of the main bottlenecks in the deployment of generated forecasts in the staffing and planning tools. In this paper, we produce and evaluate the whole forecast distribution of hourly ED arrivals, using all forecasting models, which could be used as a risk management tool for planners and decision makers. Moreover, datasets used in some of these studies are relatively small, short length (e.g. time period of 1 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 reliable. Moreover, all previous publications covered in this paper fails reproducibility principles.

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

1. We develop a novel methodology to forecast short-term hourly hospital admission using the family of Generalised Additive Models that accounts for i) ..., ii) ... iii) , ... iv) (Jethro to complete)
2. We produce probabilistic forecast, in addition to the point estimation, that quantify uncertainties in future hospital admission and we evaluate its accuracy using ...;
3. We benchmark the accuracy of our model against appropriate models used when multiple seasonality is present, i.e. Prophet, TBATS, Poisson Regression, and exponential smoothing state space model (ETS);
4. We provide data and code enabling reproduction and refinement of the proposed approach and benchmarks. The proposed approach could also be generalized to forecast hourly requirements in other services such as the number of incidents or call volumes in clinical desk services.

Our contributions are as following:

The rest of the paper is organised as following: section 2 provides a brief overview of hourly forecasting in the healthcare Section 4 starts with ...

2. Research background: hourly ED forecasting

There are many studies that employs models to forecast the Emergency Department needs such as admission and arrivals. The time granularity considered by these studies spans from hourly to yearly. However, given the focus of the paper, we only focus on studies on hourly ED forecasting. Table 1 summarize studies in hourly forecasting in the emergency department.

Table 1: Summary of studies in hourly Emergency Department forecasting

Author	Year	Variable	Horizon	Length	Method	Metric	Probabilistic	Multiple seasonality
McCarthy 2008	2008	ED arrivals	24h	2 years	Poisson log-linear regression model	95% CI	No	Yes
Asheim et al.	2019	ED arrivals	3h	5	Poisson regression	MAPE	No	No
Cote et al.	2013	ED arrivals	24h	2	Fourier regression	R^2, Standard Error	No	No
Kim et al.	2014	Hospital admission	4h, 24h	3	Linear regression; Exponential smoothing; ARIMA; GARCH; VAR	MAPE	No	No
Schweigler et al.	2009	Bed occupancy	4h, 12h	4	Hourly historical average; SARIMA; Sinusoidal model with autocorrelated error	RMSE	No	No
Channouf et al.	2007	Ambulance admission	1h, 3h, 6h, 12h, 13h, 14h, 17h, 23h, 24h	2	Regression	RMSE	No	No
Hertzum	2017	ED arrivals	1,2,4,8,24 hours	3	linear regression; SARIMA; Naïve	MAE, MAPE, MASE	No	No
Choudhury and Urena	2020	ED arrivals	1h to 24h	4	ARIMA; Holt-winters; TBATS; ANN	RMSE, ME	No	Yes
Jones et al.	2009	ED census	24h	2	VAR; Holt winters	MAE	No	No
Morzuch and Allen	2006	ED arrivals	168h	3	Regression; ARIMA; Exponential smoothing	RMSE	No	No
Chase et al.	2012	ED CUR	30m 1h, 2h, 4h, 8h, 12h	1	Binary regression	NA	No	No

Linear regression, ARIMA, and naive models were used by Hertzum [11] to investigate whether accurate hourly accident and emergency department patient arrivals and occupancy forecasts can be generated using calendar variables. Naive model was there for the purpose of comparison. Hertzum [11] study shows that patient arrivals variation is larger across the hours of the day than across the days of the week and the months of the year. In term 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 days. July-August are the month with the highest number of patient arrivals and January and February are the months with the lowest number of arrivals. The regression and ARIMA models perform similarly for all forecast interval in modeling patient arrivals. In modeling accident and emergency department occupancy, ARIMA outperform regression models. However, after all, the models of occupancy were less accurate than those arrivals. Hertzum [11] mentioned that ARIMA models are among the most accurate models for accident and emergency department visits forecasting. Another interesting point is that the accuracy of accident and emergency department forecasting models decrease with the increasing forecast interval. Lastly, the accuracy of the forecasting model may possibly be increased with additional information added to the model.

Predicting the arrivals of an ED future arrivals is also studied by Choudhury and Urena [6]. ARIMA, Holt-Winters, TBATS, and neutral network methods were implemented 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. Hertzum [11] work was mentioned in this paper. This result is surprising given the existence of multiple seasonality in the hourly dataset.

Morzuch and Allen [17] used the Unobserved Components Model (UCM), by which each component of the time series is separately modeled as stochastic. Double-seasonal exponential smoothing and standard Holt-Winters are 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. [16] employed a Poisson log-linear regression model, including independent variables such as temporal factors (i.e., 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 forecast patient demand for ED services. Hourly arrival data of ED arrivals in the 1-year study period was deployed to forecast from 1 hour to 24 hours into the future. Authors . They present the prediction interval accuracy of 50% and 90% intervals for the number of hourly arrivals under the Poisson assumptions. They show that the most important predictor is hour of the day and autocorrelation lag 1. However, these findings are limited to the short number of observations (only one year of historical data).

Schweigler et al. [23] conducted an investigation on whether using 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. In particular, 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 model 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. [14] compared different univariate and multivariate time series forecasting techniques to forecast patient volume for a Hospital Medicine programme. The study adopted historical mean linear regression as benchmark, exponential smoothing, ARIMA, SARIMA, GARCH (generalized autoregressive conditional heteroskedasticity method), able to adjust changes in variance over time and vector autoregressive (VAR) method, able to incorporate data from different sources, to forecast for 4 hours, 24 hours. They used MAPE to report the forecast accuracy. Each of the forecasting models outperformed the benchmark model. In particular, ARIMA model performed best.

Gijo and Balakrishna [10] generated a time series model to forecast the daily and hourly call volume

at all centre handling emergency ambulance services. Since historical data showed seasonality, SARIMA models were investigated. Regarding the daily model, the authors generated a SARIMA model, which, however, resulted in the forecast error (MMSE) that significantly increased when the lead time exceeded 8 days. On the other hand, the SARIMA model proposed to forecast the log-calls an hourly basis. This model was found to fit well the model both for shorter and longer lead times.

Asheim et al. [1] developed a Poisson time-series regression model with continuous weekly and yearly cyclic effects to implement a real-time system that could forecast ED arrivals on 1,2,3 hours horizon. Once measured the accuracy using the MAPE metric, it was noticed that great improvement happened when time of notification was incorporated into the model, especially on a one-hour horizon. Therefore, time of patient notification must be available for this model to be successful.

According to the studies mentioned earlier, it can be said that the existing studies 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 said that the accuracy of hourly accident and emergency department forecasting model is low compared to other longer forecasting intervals like daily forecast [3, 11]. However, some studies mentioned that the accuracy of ED hourly forecast is at the acceptable level [6, 16, 23].

There are few limitations in the literature which encourage us to undertake this research and propose examine different forecasting approaches to deal with them. 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 forecast and consequently report only point forecast accuracy. There is a lack of studies presenting probabilistic forecast of hourly ED arrivals that better represent uncertainty of future admissions, providing an holistic picture of future demand for a planner; (iii) most studies are not reproducible, as it is almost impossible to reproduce forecasting models and results from the studies; (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 nor is forecast accuracy reported.

3. Experimental design

3.1. data

Data used in the study comprised 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 is used for empirical evaluation 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 but it reveals some systematic structures.

It is clear that the number of arrivals has a sub-daily structure. The ED arrivals decreases between mid-night and early morning and then increases until the evening and then decreases again. It is also clear that ED service gets systematically more visits on Monday between 8 a.m. and 5 p.m. Moreover, The number of arrivals around mid-night is slightly higher for Saturday and Sunday.

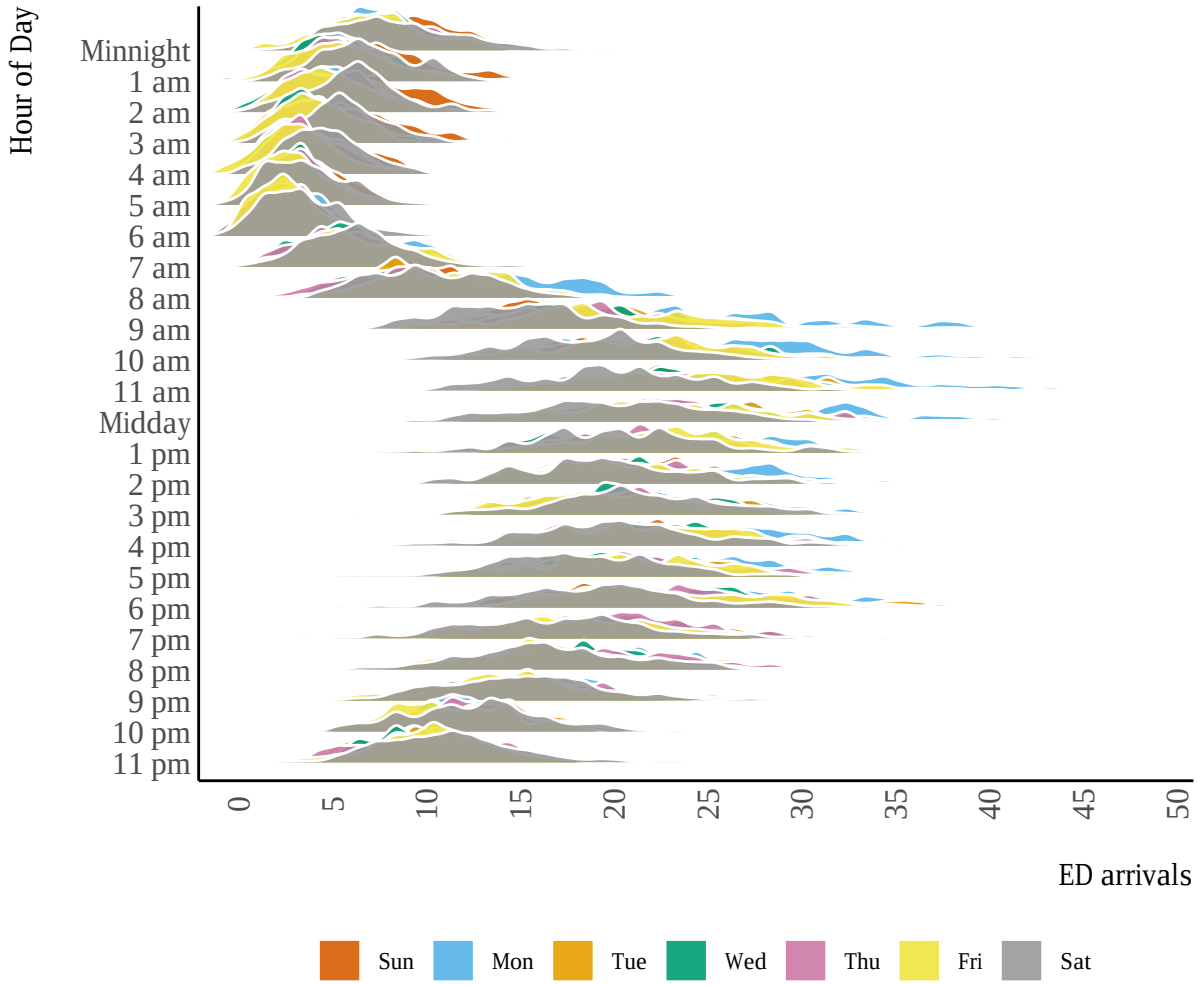

Figure 1: Distrubution of admission per hour and day of the week

Figure 2 also highlights that there are many ouliers for almost every hour of the day that may affec the accuracy of forecasting methods, some of these outliers are might be related to holidays and special events. It is also clear that arrivals are less volatile between mid-night and early morning.

We propose a forecasting model that accounts for the systematic structure of the data.

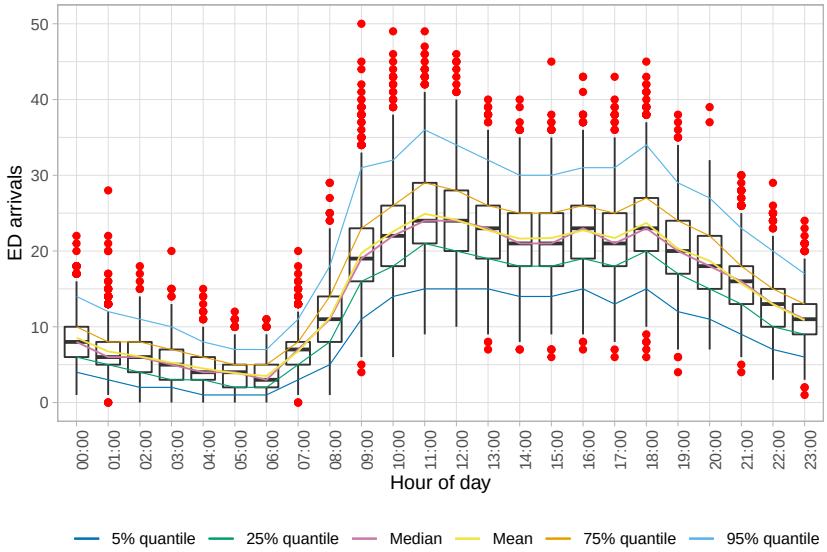

Figure 2: Seasonal plot of ED attendance

3.2. Benchmarks

3.2.1. Naive/Climatology

Empirical distribution for same hour-of-day and day-of-week as target time.

3.2.2. Multiple Regression

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

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

where $\mathbf{x}_t$ is the vector of explanatory variables, $\boldsymbol{\beta}$ is the vector of parameters and ϵ_t is the error term, which is typically assumed to follow Normal distribution with zero mean and a fixed variance. However, in the context of healthcare and A&E admittance, 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, which can be summarised as:

$$\mathbf{y}_t \sim \text{Poisson}\left(\exp\left(\mathbf{x}_t'\boldsymbol{\beta}\right)\right). \tag{2}$$

The exponent 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 one correct answer and the decision needs to be done based on 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` v0.7.0 package [24] for R [31] for the experiments and denote this model as "Poisson Regression".

3.2.3. ETS - Exponential Smoothing model

Hyndman et al. [13] 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, 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 [25] 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.

3.2.4. Prophet

Prophet is a forecasting procedure created by Facebook [30] 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 popular and automated, making it easy to learn for beginners. The implementation may be less flexible than other methods. The model is incorporated using corresponding implementation of the Fable package in R. We use the `prophet()` function in the fable package to generate hourly forecasts [19]. Note that but the input data is assigned with an hourly and daily seasonality.

3.2.5. TBATS

De Livera et al. [7] proposed a model to deals with time series exhibiting multiple complex seasonalities. TBATS includes a Box-Cox Transformation, ARMA model for residuals and a trigonometric expression of seasonality terms. The later one not only gives the model more flexibility to deal with complex seasonality but also reduces the parameters of model when the frequencies of seaonalities are high. We fit a TBATS model to our daily time series using the `tbats()` function in the forecast package of R [12].

4. Model building

4.1. ADAM: multiple seasonal iETSX

Svetunkov [26] proposed a framework for dynamic models called **ADAM** - Augmented Dynamic Adaptive Model. This framework includes ARIMA [2], ETS [13] and regression, supporting multiple frequencies, non-normal distributions and intermittent demand [27]. 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. Given that the data exhibits zeroes, we use the direct probability of ETS(M,N,N) model developed by Svetunkov and Boylan [27] 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 $(1+\epsilon_t) \sim -(s^{-1}, s)$, 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. We expect this model to perform better than Poisson regression, because it has dynamic parts (level and seasonal components), but also takes external information into account. 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. This model is implemented in `adam()` function from `smooth` package [25] for R and is denoted in our experiment as "ADAM".

4.2. GAMLSS

If we assume that our predictive distribution follows a given parametric distribution, 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 of the assumed distribution may be modelled as additive functions of explanatory variables. This provides a powerful and flexible framework for probabilistic forecasting, provided that a suitable distribution and additive structures can be found. In practice, this means experimenting with various distributions and evaluating their suitability using available training data. Let y_t be the number of attendances in time period t and indicate with $F_t(y_t)$ its predictive cumulative probability distribution. 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 Rigby and Stasinopoulos [21] the elements $j = 1, ..., m$ of $\boldsymbol{\theta}_t$ are modelled via

If we assume that our predictive distribution follows a given parametric distribution, 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 of the assumed distribution may be modelled as additive functions of explanatory variables. This provides a powerful and flexible framework for probabilistic forecasting, provided that a suitable distribution and additive structures can be found. In practice, this means experimenting with various distributions and evaluating their suitability using available training data.

Let y_t be the number of attendances in time period t and indicate with $F_t(y_t)$ its predictive cumulative probability distribution. 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 [21] the elements $j = 1, ..., m$ of $\boldsymbol{\theta}_t$ are modelled via

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

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 $S_{j,i} \subset \{1, \ldots, d\}$. 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 element 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 **?**] 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, given by

$$F_t(y_t, \lambda_t) = \frac{\Gamma(\lfloor y_t + 1 \rfloor, \lambda_t)}{\lfloor y_t \rfloor} \tag{6}$$

where we consider an additive model for λ_t of the form

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

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.

Truncated Normal...

$$F_t(y_t, \mu_t, \sigma_t) = \frac{\Phi\left(\frac{y_t - \mu_t}{\sigma_t}\right) - \Phi\left(\frac{-y_t}{\sigma_t}\right)}{1 - \Phi\left(\frac{-y_t}{\sigma_t}\right)} \tag{8}$$

with

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

Truncated t distribution...

Negative binomial...

4.3. Forecast performance evaluation

In order to assess performance of models, we track quantiles (5th, 10th, etc up to 95th quantile) and conditional expectations for 48 steps ahead for each model. The forecasts are produced every 12 hours for the holdout of 365 days in a rolling origin fashion [29], 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):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} e_j^2}. \tag{9}$$

The objective of density forecasts it to be as sharp as possible while remaining reliable/calibrated [**?**]. A forecast is said to be sharp if the predictive distribution has a relatively small spread, indicating low uncertainty, witch is valuable to decision makers provided the forecast is calibrated.

The Pinball Score is a strictly proper score used to evaluate quantile forecasts and is the discrete form of the Continuous Rank Probability Score. 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 estimation. The Pinball Score is given by

$$\mathrm{Pinball} = \frac{1}{T|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \sum_{t=1}^{T} (q_{\alpha,t} - y_t) \left(\mathbb{1}(y_t \leq_{\alpha,t}) - \alpha\right) \tag{10}$$

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

Calibration, also called `reliability`, is the property that forecast probabilities match the observed frequency of realisations. If a forecast is calibrated, then 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 observer frequency mean($\mathbb{1}(y_t \leq q_{\alpha,t})$).

To compare model performance, and the significance of any apparent difference in performance, it is useful to define skill scores. Skill scores may be calculated for any metric using

$$\text{Skill} = \frac{M_{\text{ref}} - M}{M_{\text{ref}} - M_{\text{perf}}} \tag{11}$$

where M is the metric's value for the method being considered, M_{ref} is the metric's value for a reference method, and M_{perf} is the metrics value for the 'perfect' method, which is zero in the case of RMSE and Pinball. 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 [9]. Here we use the best performing simple benchmark, Benchmark-2, as the reference model, and employ a block-boostrap with blocks of length 24h in order to account for temporal correlation of the underlying data.

5. Result and discussion

(Plots generated with theme_few - they look nice, good suggestion! Not sure how to include them in bookdown, can you help, Bahman? The plots are in the results folder for now, can either point to them from here, or I can change the Evauation script to save them somewhere else, and maybe in a different format? I guess PDF might be preferred. I realise that while I prepared my secript to produce point forecasts I haven't run it yet so will do this next week.)

5.1. Case study

The data is portioned into training and test data, with all model development and hyper-parameter tuning performed using training data only. Data from 2014-04-01 to 2018-02-28 are used for training, and from 2018-03-01 to 2019-02-28 for testing.

(Summary of data, e.g. discussion of observations from training data/model development: long-term trend)

5.2. Forecast evaluation

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 **??**, 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 iETS 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.

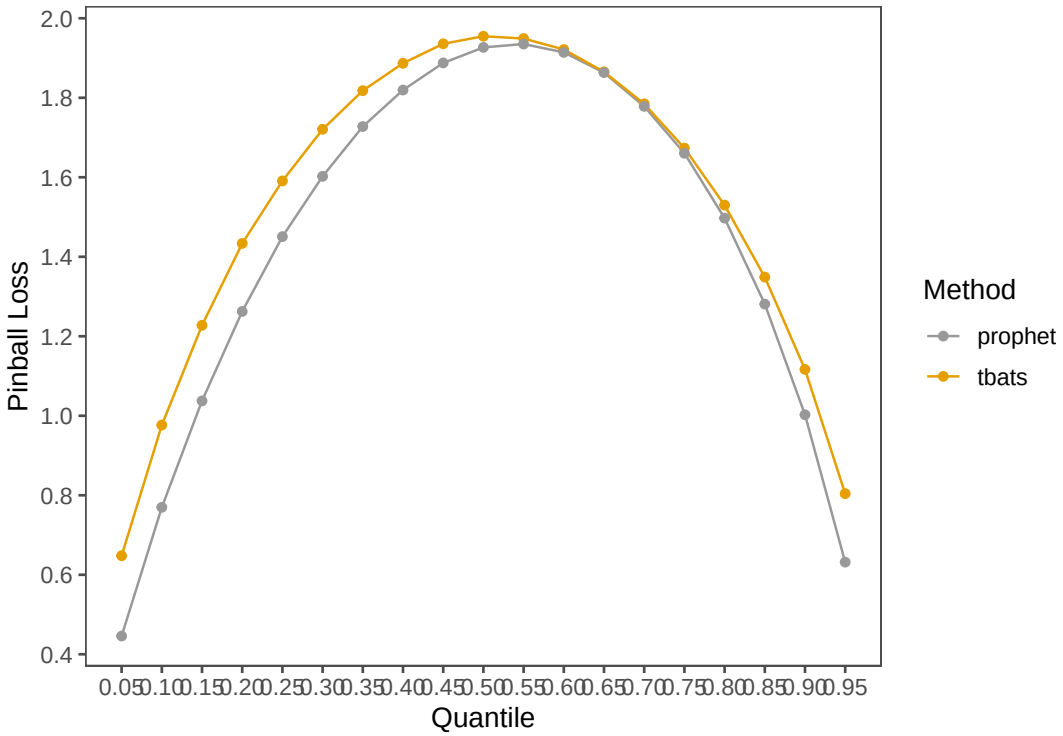

Figure 3: pinball ...

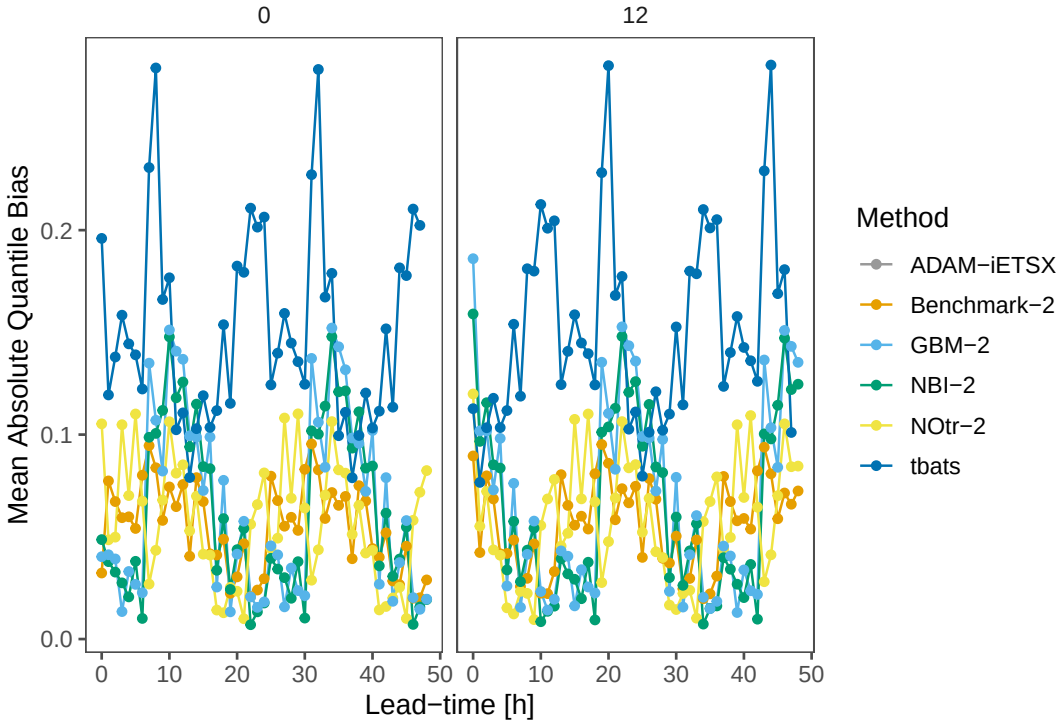

Figure 4: pinball ...

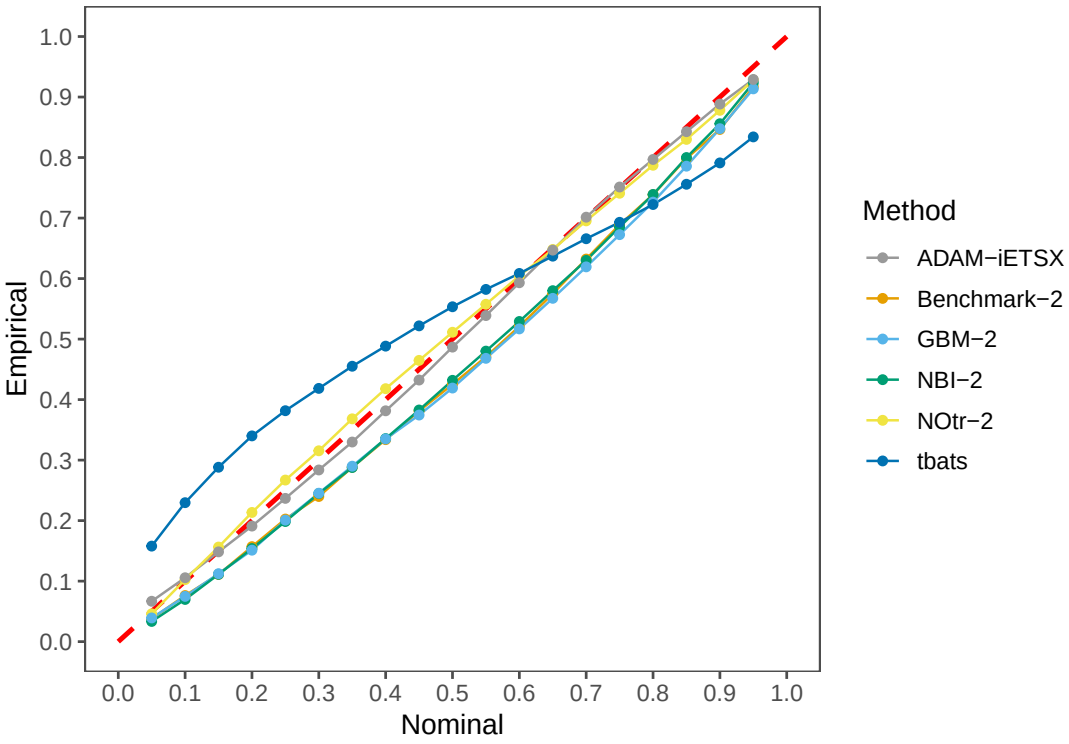

Figure 5: Reliability ...

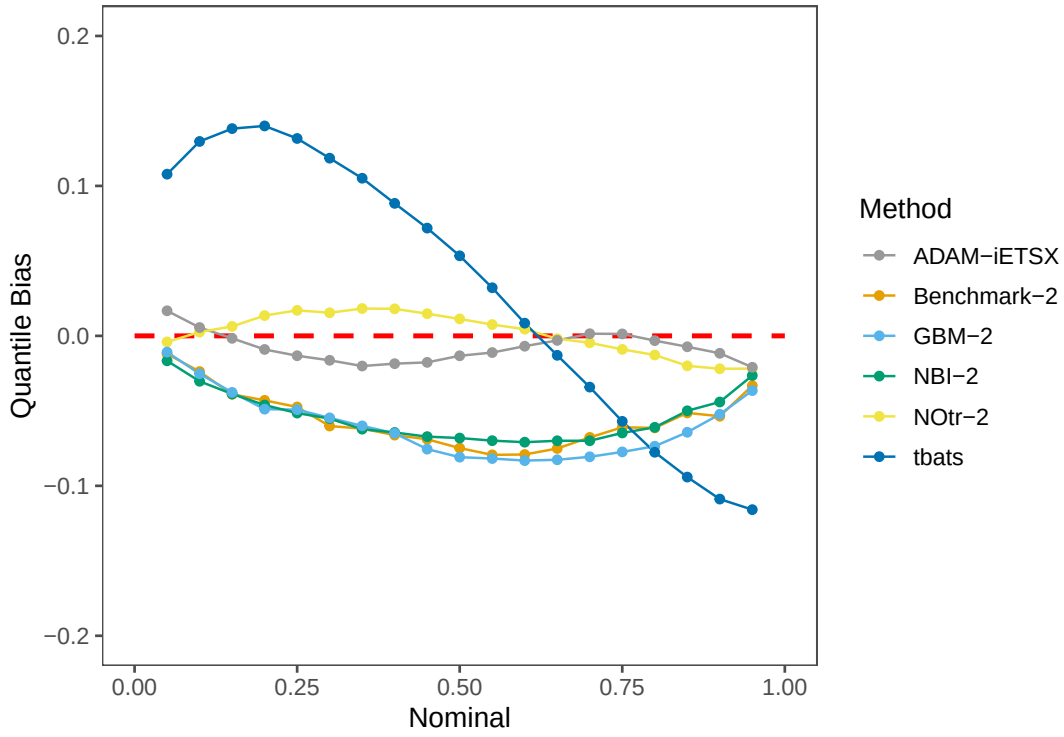

Figure 6: Quantile ...

Evaluation metrics from the test period are presented in **??** which are ordered by Quantile Bias. The five models identified above have a Quantile Bias of 0.014 or less, which is substantially lower

Table 2: Summary of studies in hourly emergency care forecasting

Method	Quantile Bias	Pinball	RMSE
Benchmark-1	0.1047874	1.254491	NA
Benchmark-2	0.0557392	1.217429	NA
ETS(XXX)	0.0194389	1.434862	0.0121247
GBM	0.0600650	1.263135	NA
NBI-2-I	0.0540740	1.206238	NA
NBI-2-log	0.0375705	1.188041	NA
NOtr-1	0.0098936	1.222584	NA
NOtr-2	0.0118522	1.208556	NA
Poisson-1	0.0372137	1.204936	NA
Poisson-2	0.0373884	1.188107	NA
Poisson-2-I	0.0523937	1.222114	NA
RegressionPoisson	0.0929416	1.293524	0.8490258
Ttr-2	0.0140252	1.210105	NA
faster	0.1862190	35.051012	NA
iETSCeiling	0.0104673	1.419490	NA
iETSXSeasonal	0.0104673	1.417260	0.0896228
qreg-1	0.0643451	1.340557	NA
tbats	0.0855702	1.536080	0.4859770

than the next group of forecast with Quantile Biases of 0.037 and above, ETS(XXX) being the only exception with a value of 0.019.

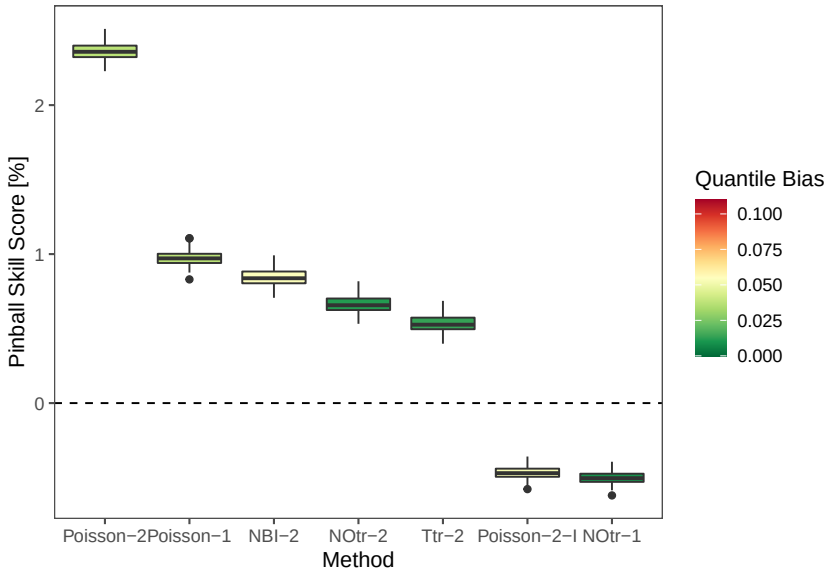

Figure 7: Skill score ...

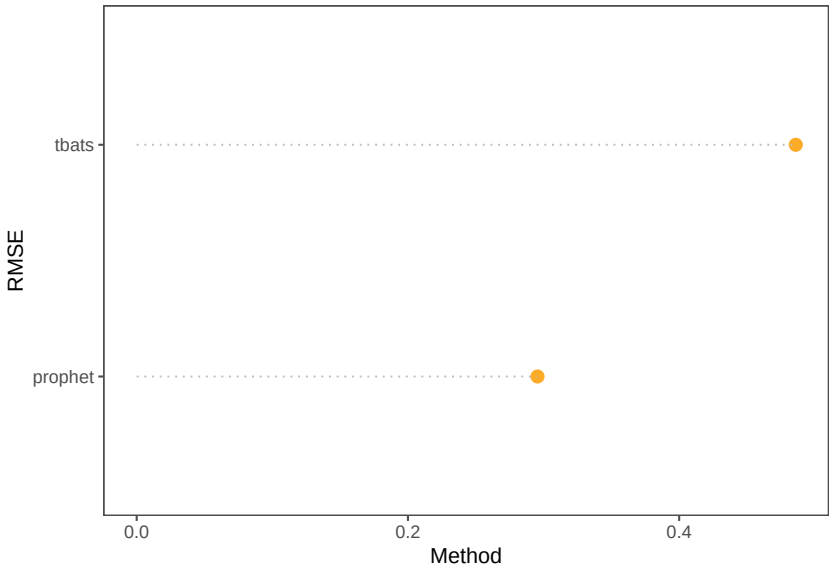

Figure 8: RMSE ...

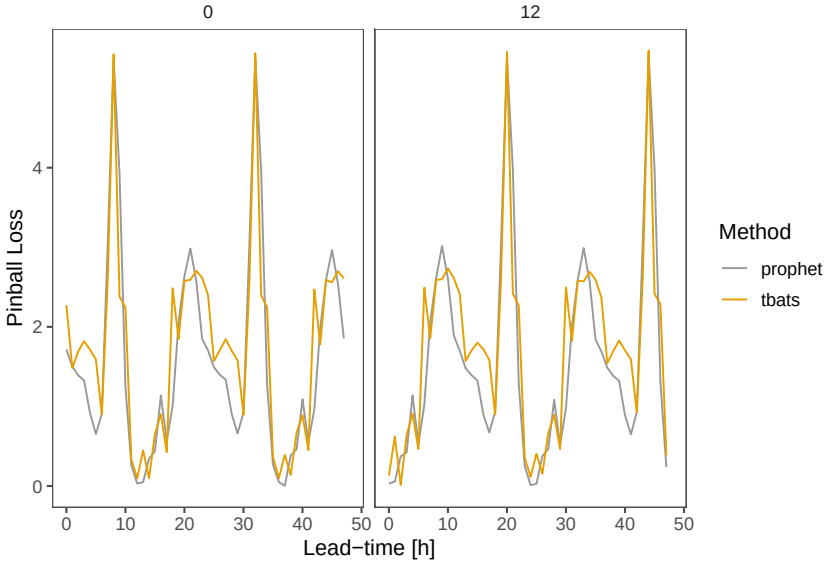

Figure 9: RMSE ...

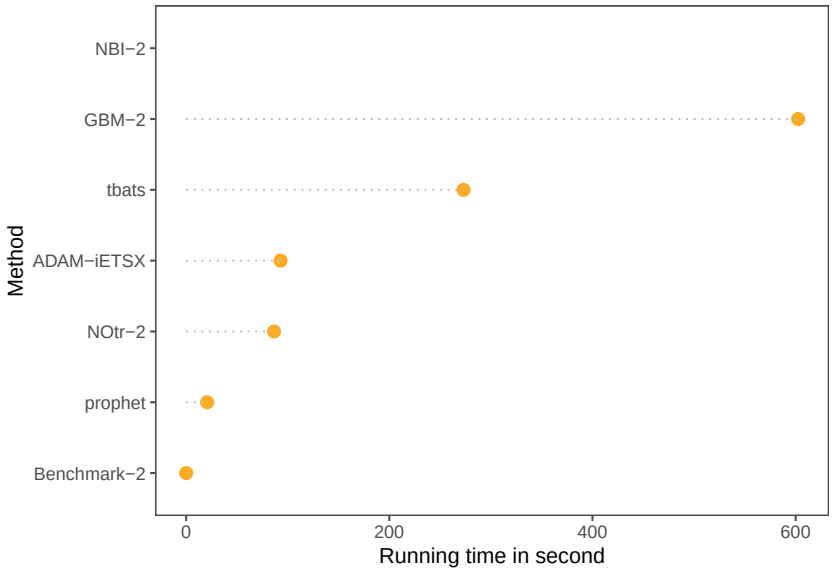

Figure 10: Running time ...

6. Conclusion

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

[3] 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.

[4] Chieh-Fan Chen, Wen-Hsien Ho, Huei-Yin Chou, Shu-Mei Yang, I-Te Chen, and Hon-Yi Shi. Long-term prediction of emergency department revenue and visitor volume using autoregressive integrated moving average model. *Computational and mathematical methods in medicine*, 2011, 2011.

[5] 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.

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

[7] 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.

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

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

[10] EV Gijo and N Balakrishna. Sarima models for forecasting call volume in emergency services. *International Journal of Business Excellence*, 10(4):545–561, 2016.

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

[12] 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.

[13] 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.

[14] 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.

[15] Qun Mai, Patrick Aboagye-Sarfo, Frank M Sanfilippo, David B Preen, and Daniel M Fatovich. Predicting the number of emergency department presentations in w estern a ustralia: A population-based time series analysis. *Emergency Medicine Australasia*, 27(1):16–21, 2015.

[16] 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.

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

[18] 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.

[19] 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.

[20] JC Park, BP Chang, and N Mok. Time series analysis and forecasting daily emergency department visits utilizing facebook's prophet method. *Annals of Emergency Medicine*, 74(4):S57, 2019.

[21] 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.

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

[23] 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.

[24] Ivan Svetunkov. *greybox: Toolbox for Model Building and Forecasting*, 2021. URL https://github.com/config-i1/greybox. R package version 0.7.0.

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

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

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

[28] Dan Tandberg and Clifford Qualls. Time series forecasts of emergency department patient volume, length of stay, and acuity. *Annals of emergency medicine*, 23(2):299–306, 1994.

[29] 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.

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

[31] 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/.

[32] Julie Leanne Vile, Jonathan William Gillard, Paul Robert Harper, and Vincent Anthony Knight. Time-dependent stochastic methods for managing and scheduling emergency medical services. *Operations Research for health care*, 8:42–52, 2016.