Models for forecasting time series data with multiple seasonalities

A research proposal submitted for the degree of

Bachelor of Commerce (Hons)

by

Mitchell O’Hara-Wild

25161407

Department of Econometrics and Business Statistics

Monash University

Australia

May 2017

Contents

1 Introduction & Background Information **1**

2 Literature Review **4**

2.1 Regression models 4

2.2 State space models 8

2.3 Benchmark model 11

3 Research plan **12**

3.1 Model evaluation methodology 12

3.2 Data 15

3.3 Proposed model 18

3.4 Implementation 19

4 Preliminary Results **20**

4.1 Model flexibility 20

4.2 Model accuracy 21

4.3 Model speed 21

A Spline interactions with seasonality **22**

B Prophet model growth equations **23**

C Holt-Winters structural form **24**

Bibliography **25**

Chapter 1

Introduction & Background Information

The availability of high frequency time series data is growing rapidly. Sensors collecting this data are ubiquitous in systems including utilities, consumer electronics, transportation, online services, and health systems. Effective use of this data plays an important role in understanding these complex systems, and accurately modelling their patterns for the purpose of forecasting. The high frequency nature of these time series often reveals multiple seasonal patterns, which require more sophisticated and flexible approaches to model them.

For example, figure 1.1 shows the complexity introduced when observing pedestrian traffic at higher frequency. When observing pedestrian counts at a daily interval (panel 1), the sub-daily patterns featuring strong seasonality during the typical working hours (panel 2) are lost. The sub-sample in figure 1.1 also shows the effects of the Easter holiday, in which the traffic pattern on public holidays are similar to that of a typical weekend. Being able to effectively use this added information when modelling should provide more accurate and useful forecasts.

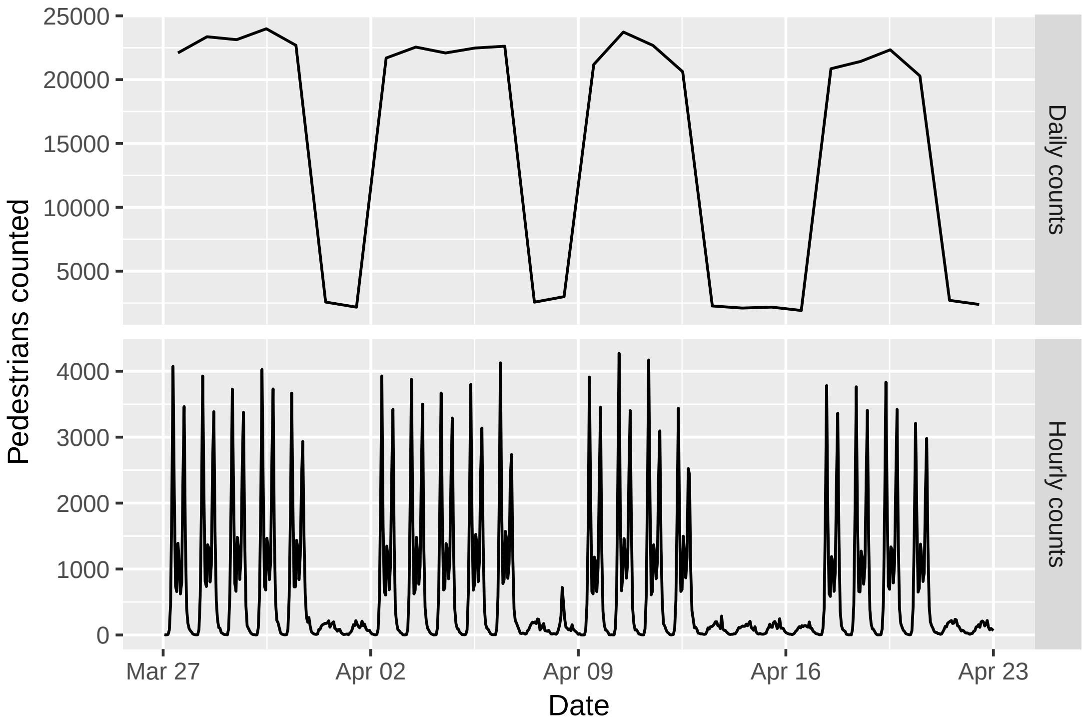

Figure 1.1: *Pedestrian traffic at Southern Cross Station featuring holiday effects*

Models for forecasting multiple seasonality time series data are predominantly built on either regression (section 2.1) or state space (section 2.2) frameworks. In the literature review (chapter 2), three commonly used models from each framework will be explored mathematically to identify how they handle patterns of multiple seasonality, among other time series features. Many of these models are methods of using more general and flexible model types, in which case a common specification for using these models for multiple seasonal forecasting will be discussed. Flexibility is common among multiple seasonal models, as it is necessary to accurately include the complex patterns in the time series data. The available literature on multiple seasonal time series modelling is sparse, so this project's literature review aims to summarise and consolidate commonly used techniques in order to compare their structure and efficacy. Throughout the research, these models will be tested on various publicly available datasets which are relevant to many multiple seasonality forecasting problems (section 3.2). As detailed in section 3.1.2, these tests will be evaluated based on forecast accuracy and computational speed, as the ideal model is highly accurate and has minimal computational cost (Gould et al. 2008).

The preliminary results outlined in chapter 4 indicate that the accuracy of the commonly used models can be improved. The regression based methods are flexible in their specification allowing a reduction in parameterisation complexity, making them relatively quick to compute. However their shortcoming is that the model parameters for trend and seasonality are inflexible with time, they typically do not handle serial correlation without ARMA errors (Shumway and Stoffer 2011), and the parameterisation can quickly become complex. The state space based approaches do not suffer from this issue, although they are prohibitively slow to estimate, and are much less flexible for including holiday effects and covariates. The research area of multiple seasonality time series forecasting is evidently in its infancy, and currently used methods can be improved. This has motivated the development of a new model in the state space framework with improved flexibility.

This research project aims to contribute to the area of multiple seasonal time series forecasting by comparing the structure and performance of commonly applied models. Testing model performance on practical time series forecasting problems featuring multiple seasonality will provide a benchmark for the evaluation of a proposed state space model to be developed and implemented throughout the year. The proposed model should allow more flexible specification, allowing for the usage of exogenous inputs whilst reducing the parameterisation problem to allow for more accurate and computationally efficient modelling.

Chapter 2

Literature Review

2.1 Regression models

2.1.1 Multiple linear regression

Linear regression is a versatile approach to modelling both cross-sectional and time-series data, where the response variable is linearly dependent on some regressors that are combined additively (Shumway and Stoffer 2011). When using multiple linear regression to forecast time series, it is common to represent the seasonal components using dummy variables that cause shifts in the forecast level. Multiple seasonality can be represented using multiple sets of seasonal dummy variables. For instance, the pedestrian sensor data has a pattern for each day of the week (denoted by $\boldsymbol{d}$), and a nested hourly seasonal pattern (denoted $\boldsymbol{h}$). Using MLR, this would be mathematically given by:

$$y_t = \beta_0 + \beta_1 d_{t1} + \cdots + \beta_6 d_{t6} + \beta_7 h_{t1} + \cdots + \beta_{29} h_{t23} + \varepsilon_t$$

Where ε_t is assumed to be i.i.d. $N(0, \sigma^2)$

This model can be further extended by introducing interactions between the seasonal terms, which is appropriate for the pedestrians dataset, as the daily pattern varies for weekdays and weekends. With interactions between $\boldsymbol{d}$ and $\boldsymbol{h}$, the model would then become:

$$y_t = \beta_0 + \beta_1 d_{t1} + \cdots + \beta_6 d_{t6} + \beta_7 h_{t1} + \cdots + \beta_{29} h_{t23} + \beta_{30} d_{t1} h_{t1} + \ldots + \beta_{167} d_{t7} h_{t23} + \varepsilon_t$$

In order to implement the larger scale movements in the time series, splines can be used to smooth over the general shape of the time series. Using f_T to describe some spline for the trend of the series (as described in section 2.1.2), the model can now be written as:

$$y_t = \beta_0 + \beta_1 d_{t1} + \cdots + \beta_6 d_{t6} + \beta_7 h_{t1} + \cdots + \beta_{29} h_{t23} + \beta_{30} d_{t1} h_{t1} + \ldots + \beta_{167} d_{t7} h_{t23} + f_T(t) + \varepsilon_t$$

As discussed in appendix A, growth in multiple seasonal models of fine temporal granularity typically should also interact with the seasonal components.

2.1.2 Generalised Additive Model

An extension of the MLR model specified above is the Generalised Additive Model (GAM), which generalises MLR by allowing the response variable to be from any distribution in the exponential family (Wood 2017). Structurally GAM is an additive regression model, and so the response variable is expressed as depending linearly on the independent variables (which are typically represented with splines). Mathematically, this makes the basic form of a GAM written as:

$$g(\mathbb{E}(y_t)) = \beta_0 + \sum_{j=0}^{p} f_j(x_{tj}) + \epsilon_t$$

Where g is the link function, ϵ is some i.i.d. random error, and f_j (for $j = 0, 1, ..., P$) are splines, defined by $f(x) = \sum_{i=1}^{q} b_i(x)\beta_i$, where b is a basis function for that particular spline.

As described by Simpson (2014), seasonality can be modelled with a GAM by using splines fitted over a seasonal identifier. For instance, to include daily seasonality, we can define a spline $f_d(x_{td})$ where $x_{td} \in 0, 1, \ldots, 6$ is the day of the week. Similarly to MLR, further seasonal patterns can be introduced additively (Laurinec 2017), so we can define hourly seasonality as $f_h(x_{th})$, where $x_{th} \in 0, 1, \ldots, 23$ describes the particular hour at time t. This makes a multiple seasonal GAM model, which can be expressed as:

$$g(\mathbb{E}(y_t)) = \beta_0 + f_d(x_{td}) + f_h(x_{th}) + \epsilon_t$$

As the independent variables can be represented using splines, there are several ways in which the model terms can interact (Laurinec 2017).

- $x_d \times x_h$: An interaction as discussed in MLR
- $f_d(x_d) \times x_h$: An interaction between a spline and a variable
- $f_d(x_d) \times f_h(x_h)$: An interaction between two splines
- $f_d(x_d) \otimes f_h(x_h)$: The tensor product interaction between two splines
  i.e. $f_d(x_d) \otimes f_h(x_h) = \sum_{i=1}^{I} \sum_{j=1}^{J} \delta_{ij} b_{di}(x_d) b_{hj}(x_h)$

Appropriate choice and inclusion of seasonal interactions are essential, as hourly patterns tend to vary on the day of the week. For example, if the third interaction were used $f_d(x_d) \times f_h(x_h)$ (commonly represented by $f(x_d, x_h)$), the model would become:

$$g(\mathbb{E}(y_t)) = \beta_0 + f_d(x_{td}) + f_h(x_{th}) + f(x_d, x_h) + \epsilon_t$$

We can also introduce trend and other model terms in a similar way to MLR.

$$g(\mathbb{E}(y_t)) = \beta_0 + f_d(x_{td}) + f_h(x_{th}) + f(x_d, x_h) + f_b(x_{tb}) + \epsilon_t$$

2.1.3 Prophet

The Prophet forecasting model (Taylor and Letham 2017) is a recent and popular method for forecasting complex multiple seasonal time series which is built upon an additive regression model; a type of GAM. The model aims to be computationally efficient, allowing it to be applied on many large datasets.

The model includes three main components, a growth function, periodic seasonality, and holiday dummy variables, giving the mathematical form of:

$$y(t) = g(t) + s(t) + h(t)$$

Prophet's growth component, $g(t)$, is defined by either a linear or a logistic growth model, with changepoints that can be automatically selected (appendix B). Although possible to specify this structure in MLR and GAM, the inclusion of population growth models to predict trend is noteworthy as regression model trend is usually implemented using splines.

The seasonal component is approximated using a Fourier series, where the complexity of the pattern is controlled by the number of approximation terms $2N$, and P is the frequency of the seasonal period.

$$s(t) = \sum_{n=-N}^{N} c_n e^{i\frac{2\pi nt}{P}}$$

Holiday effects are included in the model by adding dummy variables:

$$h(t) = \sum_{i=1}^{L} \kappa_i \mathbf{1}_{t \in D_i}$$

Where for a some set of L holidays, $\mathbf{1}$ is an indicator function that causes a shift by κ_i if the time t is in the list D_i of dates for the holiday i. This is relatively simple implementation of holiday effects, and cannot allow holidays to interact with other timeseries features, and the shift cannot vary with time.

Although within the capabilities of the Prophet model structure, the implementation does not currently support sub-daily seasonal patterns. Therefore, its performance on the evaluation datasets (section 3.2) cannot yet be tested at the desired sub-daily level. However this feature is planned for the next version of the R package, and when it is available appropriate tests can be performed.

2.2 State space models

2.2.1 Double-Seasonal Holt-Winters

An early approach to forecasting seasonality using exponential smoothing is the Holt-Winters method proposed by Holt (1957), and extended by Winters (1960). This method has since been written in a variety of forms, including structural form (appendix C), and in the innovation state space model form Ord, Koehler, and Snyder (1997) below:

$$
\begin{aligned}
y_t &= \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t \\
\ell_t &= \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t \\
b_t &= b_{t-1} + \beta\varepsilon_t \\
s_t &= s_{t-m} + \gamma_m\varepsilon_t
\end{aligned}
$$

Taylor (2003) extended this single seasonal model by including a second seasonality term, which is known as the Double-Seasonal Holt-Winters (DSHW) model:

$$
\begin{aligned}
y_t &= \ell_{t-1} + b_{t-1} + s^{(1)}_{t-m_1} + s^{(2)}_{t-m_2} + \varepsilon_t \\
\ell_t &= \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t \\
b_t &= b_{t-1} + \beta\varepsilon_t \\
s^{(1)}_t &= s^{(1)}_{t-m_1} + \gamma_1\varepsilon_t \\
s^{(2)}_t &= s^{(2)}_{t-m_2} + \gamma_2\varepsilon_t
\end{aligned}
$$

As this is a state space model, the seasonal components can update over time, however holiday effects are not possible in this model. Also, as it is specified as an innovations state space model, co-regressors cannot be used.

2.2.2 BATS and TBATS

The BATS model proposed by De Livera, Hyndman, and Snyder (2011) is an extension of Taylor's DSHW model to include a Box-Cox transformation, ARMA errors, trend dampening, and an unrestricted amount of seasonal patterns. In the innovations state space structure, this model is given by:

$$
\begin{aligned}
y_t^{(\lambda)} &= \begin{cases} \dfrac{y_t^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y_t) & \text{otherwise} \end{cases} \\
y_t^{(\lambda)} &= \ell_{t-1} + \phi b_{t-1} + \sum_{i=1}^{M} s_{t-m_i}^{(i)} + d_t \\
\ell_t &= \ell_{t-1} + \phi b_{t-1} + \alpha d_t \\
b_t &= (1-\phi)b + \phi b_{t-1} + \beta d_t \\
s_t^{(i)} &= s_{t-m_i}^{(i)} + \gamma_i d_t \\
d_t &= \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t
\end{aligned} \tag{2.2.1}
$$

As seen above in equation (2.2.1), the response y_t is transformed with a Box-Cox transformation with parameter λ, which gives $y_t^{(\lambda)}$, d_t is an ARMA process replacing ε_t, ϕ controls the trend dampening, and each seasonal term $s_t^{(i)}$ are additively included in the model.

The TBATS model further extends BATS by using a trigonometric representation of seasonal components based on a Fourier series. The model follows the same structure in equation (2.2.1), however now, we replace the seasonal terms $s_t^{(i)}$ with:

$$
\begin{aligned}
s_t^{(i)} &= \sum_{j=1}^{k_i} s_{j,t}^{(i)} \\
s_{j,t}^{(i)} &= s_{j,t-1}^{(i)} \cos \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin \lambda_j^{(i)} + \gamma_1^{(i)} d_t \\
s_{j,t}^{*(i)} &= -s_{j,t-1}^{(i)} \sin \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos \lambda_j^{(i)} + \gamma_2^{(i)} d_t
\end{aligned}
$$

$$\lambda_j = \frac{2\pi j}{m_i}$$

Where $\gamma_1^{(i)}$ and $\gamma_2^{(i)}$ are smoothing parameters, and k_i is the number of harmonics for the seasonal component i. This method of formulating seasonality splits the seasonality into the level of the seasonality ($s_{j,t}^{(i)}$), and the change in the seasonality ($s_{j,t}^{*(i)}$). This is a more flexible specification than the seasonality in BATS.

2.2.3 Multiple seasonal ARIMA

As described in Hyndman and Athanasopoulos (2014), a seasonal ARIMA model can be described by $ARIMA(p,d,q)(P,D,Q)_m$, where m is the frequency of the seasonal component. This can be represented using backshift notation (where $B^k y_t = y_{t-k}$) as:

$$(1 - \sum_{i=1}^{p} \phi_i B^i)(1 - \sum_{i=1}^{P} \Phi_i B^{i\times m})(1-B)^d (1-B)^{D\times m} y_t = (1 + \sum_{i=1}^{q} \theta_i B^i)(1 + \sum_{i=1}^{Q} \Theta_i B^{i\times m}) e_t$$

The seasonal ARIMA can be further extended to handle multiple seasonal patterns by multiplying additional seasonal terms (Box et al. 2015; Au, Ma, and Yeung 2011), forming an $ARIMA(p,d,q)(P_1,D_1,Q_1)_{m_1} \ldots (P_M,D_M,Q_M)_{m_M}$:

$$\prod_{j=0}^{M} \left\{ \left(1 - \sum_{i=1}^{P_j} \Phi_{ji} B^{i\times m_j}\right) (1-B)^{D_j \times m_j} \right\} y_t = \prod_{j=0}^{M} \left(1 + \sum_{i=1}^{Q_j} \Theta_{ji} B^{i\times m_j}\right) e_t$$

Where $P_0 = p$, $D_0 = d$, $Q_0 = q$ and $m_0 = 1$. Although possible to specify two or more seasonal periods, the parameterisation of multiple seasonal ARIMA models becomes very computationally expensive.

This model also allows the inclusion of co-regressors, which also allows for the inclusion of holiday effects. For some set of time varying regressors $\boldsymbol{X}_t$ with coefficients $\boldsymbol{\beta}$, the model becomes:

$$\prod_{j=0}^{M} \left\{ \left(1 - \sum_{i=1}^{P_j} \Phi_{ji} B^{i\times m_j}\right) (1-B)^{D_j \times m_j} \right\} (y_t - \boldsymbol{X}_t \boldsymbol{\beta}) = \prod_{j=0}^{M} \left(1 + \sum_{i=1}^{Q_j} \Theta_{ji} B^{i\times m_j}\right) e_t$$

2.3 Benchmark model

2.3.1 Seasonal naïve

A common benchmark for forecasting time series with seasonality is the seasonal naïve method, which simply involves making forecasted values equal to the value which was last observed from the same seasonal period (Hyndman and Athanasopoulos 2014). Mathematically, this can be represented as:

$$\hat{y}_{T+h} = y_{T+h-km}$$

Where m is the seasonal period (and in the multiple seasonal case, the product of all seasonal periods), and k is the number of seasonal periods h predicting past, mathematically, $k = \lfloor (h-1)/m \rfloor + 1$. This simple model does not allow for trend or any other non-seasonal features.

Chapter 3

Research plan

3.1 Model evaluation methodology

3.1.1 Model specification

Particular care is needed when evaluating the benchmark models as misspecification can result in the model performing worse than it normally would. In order to fairly compare models which require specification of parameters or other choices, we need to define a decision making process to be applied to all models. If an automated procedure is provided for model selection, the default options for the implementation will be used. Otherwise, the best model specification should be selected by minimising the appropriate information criterion or measure of fit. Lastly, if such an information criterion is not available, the decisions are to be made using an optimisation algorithm that uses a validation set. Model selection should be appropriate provided the selection technique for each model type is the same.

3.1.2 Evaluation metrics

The ideal time series model should be flexible in how it can be specified, provide accurate forecasts, well conditioned to the input series, and be feasible to estimate without extensive computational burden. These characteristics are evaluated by analysing model structure, applying models to data, and introducing simulated noise and outliers.

Flexibility of model

As explored in the literature review (chapter 2), the structure of the model imposes restrictions on how the model is used, limiting its flexibility. To allow ease of comparison for these methods, we can specify a table of commonly encountered features of time series, and determine if they can be included in the structure of the model. In particular, this table should feature the time series components that not all models support, such as: handling of serial correlation, structural breaks, holiday effects, time varying parameters, and the support of covariates.

Accuracy of forecasts

Having specified the appropriate parameterisation for each of the models, their accuracy will be tested based on varying measures of performance on the datasets mentioned in section 3.2. If computationally feasible, forecast accuracy will be evaluated on the basis of a H-step rolling-origin time series cross-validation approach. To reduce the computational cost of re-fitting models into a feasible time frame, a step size of one week will be used. The algorithm for this process is given as follows:

1. Subset the first year from the dataset
2. Using this subset of data, estimate the specified model
3. Use this model to produce forecasts for the one week beyond the end of the subset
4. Calculate and save the forecast errors given by $e_{t+h} = y_{t+h} - f_{t+h}$, where $h = 1, 2, \dots, H$ (with H being the number of observations in a week), and t is the last observation in the subset.
5. Increase the size of the subset by one week, and repeat from step two until all data is used.
6. Use these errors vectors to compute commonly used out-of-sample error metrics including:
   - $RMSE_h = \sqrt{\frac{1}{v_h} \sum_{t=1}^{v_h} (e_{t \times h+1})^2}$
   - $MAE_h = \frac{1}{v_h} \sum_{t=1}^{v_h} |e_{t \times h+1}|$
   - $MAPE_h = \frac{1}{v_h} \sum_{t=1}^{v_h} |100 \frac{e_{t \times h+1}}{y_{t \times h+1}}|$

Where v_h is the total number of forecast errors calculated h steps ahead.

Note that the error measures produced by the above algorithm are seperated by h, as the errors for different number of steps ahead will differ. We can instead plot these error measures to also identify how accurate the forecasts are for varying forecast horizons. By computing the forecast accuracy using the above algorithm on each of the 43 pedestrian count sensor locations and the electricity demand dataset, a representative indication of the forecast accuracy for these methods will be achieved.

Sensitivity and robustness

Sensitivity for each of the models will also be tested by introducing outliers via simulation at varying points from the end of the series. In a long series, these outliers should not substantially change the structure or fit of the model. Additionally, outliers occuring further from the last observation in the series should have less effect than more recent disturbances. Varying amounts of white noise can also be added to the series to test how well the models perform under varying degrees of uncertainty.

Computational burden

Whilst evaluating the accuracy of the forecasts (as in 3.1.2), various times will be recorded during the estimation and forecasting computations. For models which are specified using an automated selection procedure, the time taken to estimate the model will be recorded only for the chosen model. In order to reduce extraneous factors impacting the computational speed, all model estimation times will be estimated on the same computer with the same software environment in which the computer's operation is otherwise idle. Further, as the pedestrian dataset provides 43 different series, averaging the computation time for this data should average out other extraneous effects. Measuring estimation and forecasting speed allows us to identify the computational burden imposed by these models for varying datasets, and varying tasks.

3.2 Data

3.2.1 Melbourne pedestrian sensors

Transportation typically features multiple seasonality as people go about their daily routines, and the ability to better understand and predict traffic is valuable for urban planning, event management and business development. One aspect of transportation is pedestrian traffic, which can be observed using pedestrian counting sensors that regularly report counts over given intervals.

In a dataset accessible from the City of Melbourne's Open Data Portal, pedestrian counts in hourly intervals are available from 43 sensor locations positioned in popular streets, landmarks and stations in the city. This time series is ideal for testing multiple seasonality as these locations feature daily seasonality, weekly seasonality (driven by working days), and holiday/event effects. For instance, figure 3.1 shows a common feature of multiple seasonality time series, in which weekdays present a different daily pattern to weekends. Strong peaks in pedestrian totals coincide with the start and end of typical working shifts, and a smaller peak coincides with lunch during the middle of the day. As one might expect, this pattern does not continue throughout weekends, and so a multiple seasonal pattern is evident in this time series.

Figure 3.2 reveals that this time series also features holiday effects which follow a similar pattern to weekends. This necessitates further seasonal features, or use of additional exogenous inputs to represent each holiday period.

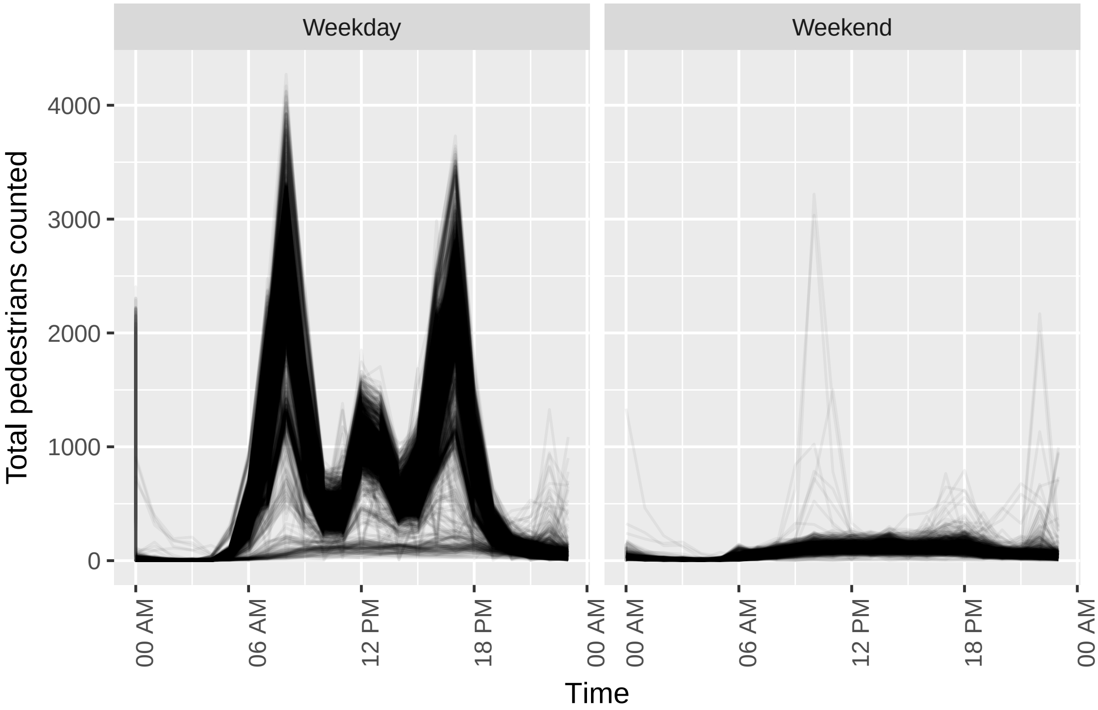

Figure 3.1: *Daily seasonality pattern in pedestrian traffic at Southern Cross Station*

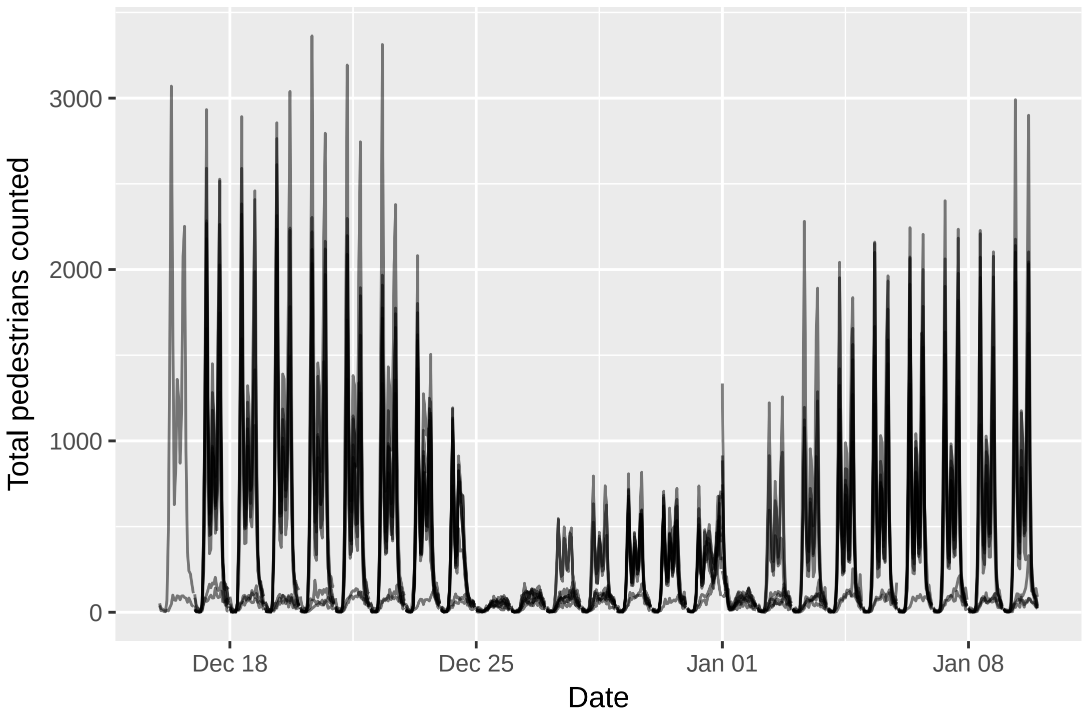

Figure 3.2: *Holiday effects in pedestrian traffic at Southern Cross Station*

3.2.2 Electricity demand for Victoria

Another time series problem which features patterns of multiple seasonality is electricity demand forecasting. Being able to accurately predict multiple seasonality in electricity demand is particularly valuable, as it would allow for more temporally fine-grained electricity load estimation. Using a dataset provided by the Australian Energy Market Operator of Victoria's total electricity demand, we can clearly see different seasonal patterns occurring on weekdays and weekends (figure 3.3). Similar to the pedestrian sensor dataset, electricity demand also features holiday effects, in which most public holidays exhibit similar electricity demand patterns to weekends.

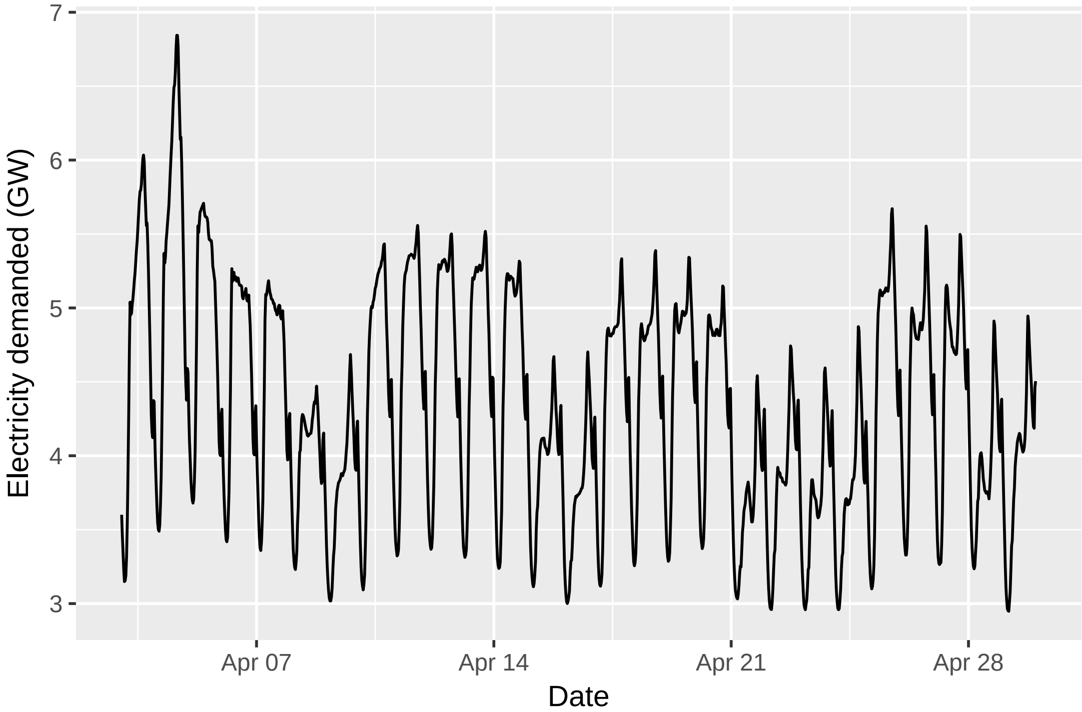

Figure 3.3: *Sub-sample of electricity demand for Victoria in 2014*

Another feature of interest in this dataset is how the pattern interacts with temperature. Using half-hourly temperatures for Melbourne (provided by the Australian Bureau of Meteorology), we can see in figure 3.4 that temperature extremes relate to an increase in electricity demand. This illustrates the importance of allowing for exogenous inputs in the proposed state-space model.

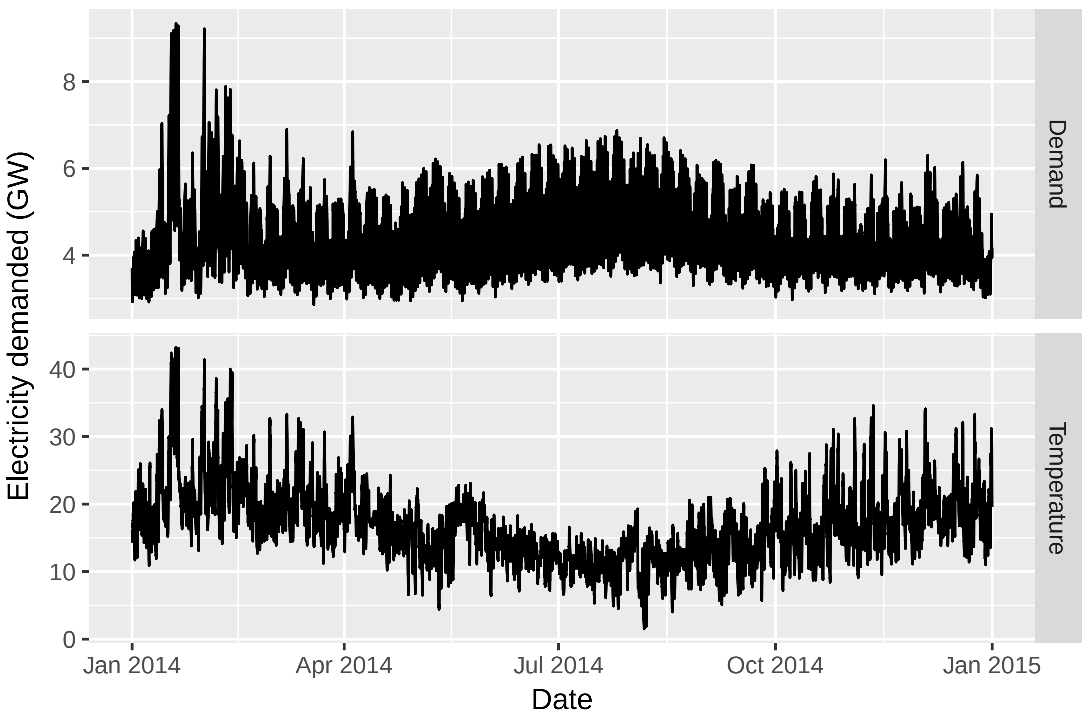

Figure 3.4: *Hourly electricity demand and temperature in Melbourne 2014*

3.3 Proposed model

Having considered the currently used methods for forecasting data with multiple seasonality, key features and improvements have been identified for the purpose of developing a new model. In particular, specifying a structural state space model is appropriate (Harvey 1990; Petris and Petrone 2011) to allow the time series parameters to update over time. This model should allow multiple sources of error to allow the inclusion of covariates in the model. Additionally, more flexible specification of seasonal terms (such as weekday and weekend instead of day of week) will reduce the computational cost by reducing the parameter space of the model.

3.4 Implementation

All analysis, writing, presentations and model creation for this research project will be performed using R (R Core Team 2013) in conjunction with many packages. In particular, the dlm package by Petris and An (2010) will be used to implement the proposed model. This project aims to adhere to the principles of reproducible research, in which the progress and history is maintained in a Git repository, and all data, code, and packages necessary to reproduce the thesis are maintained in an R package.

Chapter 4

Preliminary Results

4.1 Model flexibility

As explored in the literature review (chapter 2), the commonly used regression based approaches allow more flexible specification of terms compared to state space models. However in order to achieve this flexibility, they often require a large (possibly infeasible) number of parameters. Regression models also have difficulties with handling serial correlation (requiring inclusion of ARMA errors), and cannot allow their time series parameters to vary with time.

A key strength of state space models is that their parameters update over time, and serial correlation is typically handled by the model structure. However the commonly used methods are implemented in the innovations state space framework, which has a single source of error, and so covariates cannot be included in the model. This issue can be overcome by implementing the model with multiple sources of error (as done with multiple seasonal ARIMA).

4.2 Model accuracy

The algorithm specified in section 3.1.2 has been implemented, however most methods are prohibitively slow to run the algorithm on all pedestrian sensor locations with re-estimating the models for each successive week. Based on preliminary experimentation on the pedestrian dataset, it seems that the ARIMA model is able to get the most accurate results, however MLR and GAM may provide more accurate results when the optimal model specification is identified. DSHW, BATS, and TBATS provided unusual and inaccurate forecasts, whilst Prophet's additive growth component was unsuitable and resulted in poor forecasts (refer to appendix A).

4.3 Model speed

Generally regression based methods outperformed state space models for speed, especially when given long time series. All regression models could be estimated in a practical amount of time, however their speed was highly sensitive to the model specification (and so the optimal model may be slow). State space models were prohibitively slow on the pedestrian dataset, with the multiple seasonal ARIMA model being computationally infeasible when using the automated model selection procedure.

Appendix A

Spline interactions with seasonality

When observing time series at a finer temporal granularity, growth patterns for trend are typically exhibited only during certain timeframes, and not uniformly across all seasonal periods. As illustrated in A.1, the trend of the timeseries at an hourly level is not additive, as the minimum number of pedestrian counts does not exhibit trend. Instead, it is the spread of the seasonality which changes, and it is the maximumal values each day which is growing. If the series were to be aggregated up to the weekly level, the usual pattern of additive trend is realised, as the seasonality is smoothed over. Therefore, when considering including trend in the models with multiple seasonality, interactions should be used.

Figure A.1: *Growth in pedestrian traffic at Southern Cross Station*

Appendix B

Prophet model growth equations

The growth component of Prophet's model with changepoints is defined by:

$$g(t) = (k + \boldsymbol{a}(t)^T \boldsymbol{\delta})t + (b + \boldsymbol{a}(t)^T \boldsymbol{\gamma})$$

Where the parameters are defined by:

- k: The initial growth rate before adjustments
- $\boldsymbol{a}(t)$: A vector of changepoint step functions where:

  $a_j = 1 \forall t > s_j$ for a given changepoint time s_j
- $\boldsymbol{\delta}$: A vector of growth rate adjustments
- b: An offset to connect segment endpoints
- $\boldsymbol{\gamma}$: A vector to make the function continuous, where $\gamma_j = -s_j \delta_j$:

The piecewise logistic growth model is given by:

$$g(t) = \frac{C(t)}{1 + \exp(-(k + \boldsymbol{a}(t)^T \boldsymbol{\delta})(t - (b + \boldsymbol{a}(t)^T \boldsymbol{\gamma}))}$$

Where $C(t)$ is the carrying capacity of the system at time t, and $\boldsymbol{\gamma}$ is now computed using $\gamma_j = \left(s_j - b - \sum_{i<j} \gamma_i\right)\left(1 - \frac{k + \sum_{i<j} \delta_i}{k + \sum_{i \leq j} \delta_i}\right)$

Appendix C

Holt-Winters structural form

Hyndman et al. (2008) provides the following formulation of the Holt-Winters model in structural form:

$$
\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + b_t h + s_{t-m+h_m^+} \\
\ell_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \\
s_t &= \gamma(t_y - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}
\end{aligned}
$$

Bibliography

Au, ST, GQ Ma, and SN Yeung (2011). Automatic Forecasting of Double Seasonal Time Series with Applications on Mobility Network Traffic Prediction. In: *2011 Joint Statistical Meetings; July*.

Box, GEP, GM Jenkins, GC Reinsel, and GM Ljung (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

De Livera, AM, RJ Hyndman, and RD Snyder (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association* **106**(496), 1513–1527.

Gould, PG, AB Koehler, JK Ord, RD Snyder, RJ Hyndman, and F Vahid-Araghi (2008). Forecasting time series with multiple seasonal patterns. *European Journal of Operational Research* **191**(1), 207–222.

Harvey, AC (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Holt, CE (1957). *Forecasting trends and seasonals by exponentially weighted averages*. O.N.R. Memorandum 52/1957. Carnegie Institute of Technology.

Hyndman, RJ and G Athanasopoulos (2014). *Forecasting: principles and practice*. OTexts.

Hyndman, R, AB Koehler, JK Ord, and RD Snyder (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Science & Business Media.

Laurinec, P (2017). *Doing magic and analyzing seasonal time series with GAM (Generalized Additive Model) in R*. https://petolau.github.io/Analyzing-double-seasonal-time-series-with-GAM-in-R/.

Ord, JK, A Koehler, and RD Snyder (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association* **92**(440), 1621–1629.

Petris, G and R An (2010). An R package for dynamic linear models. *Journal of Statistical Software* **36**(12), 1–16.

Petris, G, S Petrone, et al. (2011). State space models in R. *Journal of Statistical Software* **41**(4), 1–25.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Shumway, RH and DS Stoffer (2011). *Time Series Analysis and Its Applications*. Springer.

Simpson, G (2014). Modelling seasonal data with GAMs. *From the Bottom of the Heap*.

Taylor, JW (2003). Short-Term Electricity Demand Forecasting Using Double Seasonal Exponential Smoothing. *The Journal of the Operational Research Society* **54**(8), 799–805.

Taylor, SJ and B Letham (2017). Forecasting at Scale.

Winters, PR (1960). Forecasting Sales by Exponentially Weighted Moving Averages. *Manage. Sci.* **6**(3), 324–342.

Wood, SN (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. 2 edition. Boca Raton: Chapman and Hall/CRC.