Sparse and Smooth Adjustments for Coherent Forecasts in Temporal Aggregation of Time Series

Souhaib Ben Taieb SOUHAIB.BENTAIEB@MONASH.EDU
Department of Econometrics and Business Statistics
Monash University
Melbourne, Australia

Editor: Oren Anava, Marco Cuturi, Azadeh Khaleghi, Vitaly Kuznetsov, Alexander Rakhlin

Abstract

Independent forecasts obtained from different temporal aggregates of a given time series may not be mutually consistent. State-of-the art forecasting methods usually apply adjustments on the individual level forecasts to satisfy the aggregation constraints. These adjustments require the estimation of the covariance between the individual forecast errors at all aggregation levels. In order to keep a maximum number of individual forecasts unaffected by estimation errors, we propose a new forecasting algorithm that provides sparse and smooth adjustments while still preserving the aggregation constraints. The algorithm computes the revised forecasts by solving a generalized lasso problem. It is shown that it not only provides accurate forecasts, but also applies a significantly smaller number of adjustments to the base forecasts in a large-scale smart meter dataset.

1. Introduction

The frequency of a given time series does not necessarily provide the best representation of it for modelling and forecasting tasks. Furthermore, rather than focusing on the given frequency, it is often crucial to generate forecasts at multiple frequencies, e.g. daily, weekly and monthly, for different decision making problems. In other words, the need is to generate forecasts for multiple *temporal aggregation* of the observed time series (see Silvestrini and Veredas (2008) for a literature review).

Forecasts for all frequencies can be computed by aggregating the forecasts of the observed time series. However, forecasting at the highest available frequency is often challenging due to the low signal-to-noise ratio, and potential useful information at other frequencies is disregarded. Instead, it is possible to independently forecast the series at all frequencies, called base forecasts. This approach allows us to use distinct methods to benefit from different levels of accuracy at various frequencies. However, using different information sets and different forecasting methods can results in incoherent forecasts, i.e. forecasts at different frequencies may not add up consistently across aggregation levels. Since the optimal forecasts are coherent by definition, it is necessary to impose coherency when generating forecasts at multiple frequencies. Also, from a decision making perspective, coherent forecasts support consistent decisions across different planning horizons.

Athanasopoulos et al. (2015) showed that temporally aggregated time series can be represented as a hierarchical time series. As a result, it is possible to use the optimal

combination framework developed by Hyndman et al. (2011) to produce coherent revised forecasts from the base forecasts. We show that the revised forecasts are obtained by adding an adjustment term to the base forecasts in order to satisfy the aggregation constraints. A similar observation has been made for contemporaneous hierarchies in Ben Taieb et al. (2017). However, with temporal hierarchies, the adjustments are applied to consecutive observations of the time series, while, with contemporaneous hierarchies, they are applied at the same time instant across time series.

If the base forecasts are unbiased, the adjustments will have a closed-form expression, and will depend on the covariance matrix of the base forecast errors (Wickramasuriya et al., 2015). We argue that the estimation errors in the error covariance matrix can lead to increased variability in the revised forecasts. We propose to compute sparse and smooth adjustments while still satisfying the aggregation constraints and minimizing forecast errors. Sparsity will allow us to keep some base forecasts unaffected by adjustments. Smoothness will provide an additional regularization by exploiting the fact that the adjustments are applied to consecutive observations of the time series. The revised forecasts are computed by solving a generalized lasso problem (Tibshirani and Taylor, 2011) for which an efficient regularization path algorithm is available (Arnold and Tibshirani, 2016). The proposed method will be compared with state-of-the-art forecasting methods using a large scale electricity smart meter data set.

2. Forecasting Temporal Hierarchies

Following the notations of Athanasopoulos et al. (2015), suppose we observe a time series $\{y_t; t = 1, \dots, T\}$ with sampling frequency m and where $T \equiv 0 \pmod m$. If $k \in \{k_p, \dots, k_1\}$ is a factor of m with $k_p = m$ and $k_1 = 1$, then the k-aggregate series with seasonal period $M_k = m/k$ can be written as $y_j^{[k]} = \sum_{t=1+(j-1)\times k}^{jk} y_t$ where $j = 1, \dots, T/k$.

To avoid using a different index j for each aggregation level k, we can define a common index i as the observation index of the most aggregated series, i.e. $y_i^{[m]}$ where $i = 1, \dots, T/m$, and an index $z = 1, \dots, M_k$ which controls the increase within each period. Then, the observations of the k-aggregate series are given by $y_{M_k(i-1)+z}^{[k]}$.

Let $\boldsymbol{y}_i^{[k]}$ be the column vector that contains the M_k observations of the k-aggregate series at time i, i.e. $\boldsymbol{y}_i^{[k]} = \left(y_{M_k(i-1)+1}^{[k]}, y_{M_k(i-1)+2}^{[k]}, \dots, y_{M_k(i-1)+M_k}^{[k]}\right)'$. Then, if we collect all these vectors in one column vector $\boldsymbol{y}_i = (y_i^{[m]}, \dots, \boldsymbol{y}_i^{[k_3]}, \boldsymbol{y}_i^{[k_2]}, \boldsymbol{y}_i^{[1]})'$, we have $\boldsymbol{y}_i = \boldsymbol{K}\boldsymbol{y}_i^{[1]}$ where $\boldsymbol{K}$ is the summing matrix given by $\boldsymbol{K} = \begin{bmatrix} \boldsymbol{I}_{k_1} \otimes \boldsymbol{1}_{M_{k_1}} \\ \vdots \\ \boldsymbol{I}_{k_p} \otimes \boldsymbol{1}_{M_{k_p}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{K}_a \\ \boldsymbol{I}_m \end{bmatrix}$, $\boldsymbol{I}_k$ is an identity matrix of order k, and $\boldsymbol{1}_{M_k}$ is a M_k-vector of ones.

In other words, the different k-aggregate series define a hierarchal time series, i.e. a multivariate time series with an hierarchical structure. Since each time series in the hierarchy is a temporal aggregation of the bottom series, we shall call it a *temporal hierarchy*.

We will denote $\boldsymbol{b}_i = \boldsymbol{y}_i^{[1]}$, a m-vector of observations at the bottom level, and $\boldsymbol{a}_i = \boldsymbol{K}_a\boldsymbol{b}_i$ a m_a-vector of observations at all aggregation levels. Then, $\boldsymbol{y}_i = (\boldsymbol{a}_i\ \boldsymbol{b}_i)'$ is a M-vector that contains all observations where $M = m_a + m$.

Let $H^* = m \times H$ be the maximum required forecast horizon for the observed series y_t, i.e. at the bottom level. Then, the required forecast horizon for each k-aggregate series is $H_k = M_k H$.

Given $I = T/m$ historical observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I$ of the temporal hierarchy, the optimal h-period ahead forecats under mean squared error (MSE) loss are given by the conditional mean (Gneiting, 2011), i.e. $\mathbb{E}[\boldsymbol{y}_{I+h}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I] = \boldsymbol{K}\ \mathbb{E}[\boldsymbol{b}_{I+h}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I]$ where $h = 1, \ldots, H$.

A natural way to compute the mean forecasts of $\boldsymbol{y}_{I+h}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I$, i.e. for all aggregation levels, is the plug-in estimator, also called bottom-up (BU), given by $\hat{\boldsymbol{y}}_{I+h} = \boldsymbol{K}\hat{\boldsymbol{b}}_{I+h}$ where $\hat{\boldsymbol{b}}_{I+h}$ is the mean forecast of $\boldsymbol{b}_{I+h}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I$.

However, the BU method does not use any series at the aggregation levels, which are often smoother and easier to forecast. A more general approach will generate forecasts for all series at all levels independently, then will compute $\hat{\boldsymbol{y}}_{I+h} = (\hat{\boldsymbol{a}}_{I+h}\ \hat{\boldsymbol{b}}_{I+h})'$ which we call the *base* forecasts.

This approach is sufficiently flexible since we can use different forecasting methods at each level. However, since the forecasts have been generated independently at each level, the aggregation constraints are not necessarily satisfied. The magnitude of constraint violation can be measured using the coherency errors, which is given by

$$\hat{\boldsymbol{r}}_{I+h} = \hat{\boldsymbol{a}}_{I+h} - \boldsymbol{K}_a\hat{\boldsymbol{b}}_{I+h}. \tag{1}$$

Definition 1 *The h-period ahead mean forecasts $(\hat{\boldsymbol{a}}_{I+h}\ \hat{\boldsymbol{b}}_{I+h})'$ are coherent if there are no coherency errors, i.e. if $\hat{\boldsymbol{r}}_{I+h} = \boldsymbol{0}$.*

3. Best Linear Unbiased Revised Forecasts

Given the possibly incoherent h-period ahead base forecast $\hat{\boldsymbol{y}}_{I+h}$, Hyndman et al. (2011) proposed to compute coherent revised forecasts $\tilde{\boldsymbol{y}}_{I+h}$ of the form

$$\tilde{\boldsymbol{y}}_{I+h} = \boldsymbol{K}\boldsymbol{P}\hat{\boldsymbol{y}}_{I+h}, \tag{2}$$

for an appropriately chosen weight matrix $\boldsymbol{P} \in \mathbb{R}^{m \times M}$.

This approach has multiple advantages: (1) the revised forecasts are generated by combining forecasts from all levels, (2) the revised forecasts are coherent by construction, and (3) multiple hierarchical forecasting methods are represented as particular cases, including the bottom-up forecasts with $\boldsymbol{P} = \left[\boldsymbol{0}_{m \times m_a}|\boldsymbol{1}_{m \times m}\right]$.

Theorem 2 *(Adapted from Wickramasuriya et al. (2015)) Let $\boldsymbol{W}_h$ be the positive definite covariance matrix of the h-period ahead base forecast errors, $\hat{\boldsymbol{e}}_{I+h} = \boldsymbol{y}_{I+h} - \hat{\boldsymbol{y}}_{I+h}$, i.e.*

$$\boldsymbol{W}_h = \mathbb{E}[\hat{\boldsymbol{e}}_{I+h}\hat{\boldsymbol{e}}'_{I+h}] = \begin{bmatrix} \boldsymbol{W}_{h,a} & \boldsymbol{W}_{h,ab} \\ \boldsymbol{W}'_{h,ab} & \boldsymbol{W}_{h,b} \end{bmatrix}. \tag{3}$$

Then, assuming unbiased base forecasts, the best (i.e. having minimum variance) linear unbiased revised forecasts are given by

$$\tilde{\boldsymbol{y}}_{I+h} = \boldsymbol{K}\tilde{\boldsymbol{b}}_{I+h}, \tag{4}$$

$$\tilde{\boldsymbol{b}}_{I+h} = \boldsymbol{P}^*\hat{\boldsymbol{y}}_{I+h}, \tag{5}$$

$$\boldsymbol{P}^* = (\boldsymbol{K}'\boldsymbol{W}_h^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}'\boldsymbol{W}_h^{-1}. \tag{6}$$

Wickramasuriya et al. (2015) and Athanasopoulos et al. (2015) used (4) to compute revised forecats for contemporaneous and temporal hierarchies, respectively. We will denote this method `MinT`.

In practice, the error covariance matrix $\boldsymbol{W}_h$ is not available, and needs to be estimated using historical observations of the base forecast errors. For contemporaneous hierarchies, Wickramasuriya et al. (2015) have estimated $\boldsymbol{W}_1$, and assumed $\boldsymbol{W}_h \propto \boldsymbol{W}_1$, since the estimation of $\boldsymbol{W}_h$ is challenging for $h > 1$. To trade off bias and estimation variance, structural assumptions on the entries of the sample covariance matrix have also been considered in Athanasopoulos et al. (2015) and Hyndman et al. (2016).

4. Sparse and Smooth Adjustments

Wickramasuriya et al. (2015) showed that the optimal weight matrix $\boldsymbol{P}^*$ given in (6) can be written as

$$\boldsymbol{P}^* = \begin{bmatrix} \boldsymbol{P}_1^* & \boldsymbol{I} - \boldsymbol{P}_1^* \boldsymbol{K}_a \end{bmatrix}, \tag{7}$$

where $\boldsymbol{P}_1^*$ depends on $\boldsymbol{K}_a$ and $\boldsymbol{W}_h$.

Now, if we plug (7) in (5), we can rewrite the `MinT` bottom revised forecasts as

$$\tilde{\boldsymbol{b}}_{I+h} = \hat{\boldsymbol{b}}_{I+h} + \boldsymbol{P}_1^* \hat{\boldsymbol{r}}_{I+h},$$

where $\hat{\boldsymbol{r}}_{I+h}$ are the coherency errors defined in (1). In other words, the `MinT` bottom revised forecasts are obtained by adding the adjustment $\boldsymbol{P}_1^* \hat{\boldsymbol{r}}_{I+h}$ to the bottom base forecasts $\hat{\boldsymbol{b}}_{I+h}$ before multiplying by matrix $\boldsymbol{K}$ to obtain the forecasts for all levels.

We propose to compute more general revised forecasts of the form $\tilde{\boldsymbol{b}}_{I+h} = \hat{\boldsymbol{b}}_{I+h} + \boldsymbol{\theta}$ where $\boldsymbol{\theta}$ is an adjustment term. This will allow us to apply structured regularization on the adjustments in order to mitigate the effect of estimation errors in $\hat{\boldsymbol{W}}_h$.

Proposition 3 *The* `MinT` *bottom revised forecasts given by* (5) *can also be computed as the Generalised Least Squares (GLS) solution of the following regression model:*

$$\hat{\boldsymbol{y}}_{I+h} = \boldsymbol{K}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{I+h}, \tag{8}$$

where $\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{b}_{I+h} | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_I]$ *and* $\boldsymbol{\varepsilon}_{I+h} \sim \mathcal{N}_M(\boldsymbol{0}, \boldsymbol{W}_h)$.

Proof The GLS solution of (8) is given by $\hat{\boldsymbol{\beta}} = (\boldsymbol{K}'\boldsymbol{W}_h^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}'\boldsymbol{W}_h^{-1}\hat{\boldsymbol{y}}_{I+h}$ which coincides with the bottom revised forecasts given by (5). ■

Given an estimate $\hat{\boldsymbol{W}}_h$ of the matrix $\boldsymbol{W}_h$, the `MinT` bottom revised forecasts can be computed as

$$\hat{\boldsymbol{\beta}}^{\text{MinT}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^m}{\arg\min} \left\{ (\hat{\boldsymbol{y}}_{I+h} - \boldsymbol{K}\boldsymbol{\beta})' \hat{\boldsymbol{W}}_h^{-1} (\hat{\boldsymbol{y}}_{I+h} - \boldsymbol{K}\boldsymbol{\beta}) \right\}. \tag{9}$$

The solution is given by $\hat{\boldsymbol{\beta}}^{\text{MinT}} = \tilde{\boldsymbol{b}}_{I+h}$ where $\tilde{\boldsymbol{b}}_{I+h}$ is defined in (5).

Now, if we apply the change of variable $\boldsymbol{\beta} = \hat{\boldsymbol{b}}_{I+h} + \boldsymbol{\theta}$ in (9), then the `MinT` adjustments are given by

$$\hat{\boldsymbol{\theta}}^{\text{MinT}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^m}{\arg\min} \left\{ (\hat{\boldsymbol{z}}_{I+h} - \boldsymbol{K}\boldsymbol{\theta})' \hat{\boldsymbol{W}}_h^{-1} (\hat{\boldsymbol{z}}_{I+h} - \boldsymbol{K}\boldsymbol{\theta}) \right\}, \tag{10}$$

where $\hat{\boldsymbol{z}}_{I+h} = \hat{\boldsymbol{y}}_{I+h} - \boldsymbol{K}\hat{\boldsymbol{b}}_{I+h} = (\hat{\boldsymbol{r}}_{I+h}\ \boldsymbol{0})'$. As discussed previously, the solution is given by $\hat{\boldsymbol{\theta}}^{\text{MinT}} = \boldsymbol{P}_1^* \hat{\boldsymbol{r}}_{I+h}$.

The amount of `MinT` adjustments $\hat{\boldsymbol{\theta}}^{\text{MinT}}$ depends on the magnitude of the consistency errors $\hat{\boldsymbol{r}}_{I+h}$, and the matrix $\hat{\boldsymbol{W}}_h$. Furthermore, since $\hat{\boldsymbol{\theta}}^{\text{MinT}}$ is a dense vector, the adjustments will affect all entries of the vector $\hat{\boldsymbol{b}}_{I+h}$. In order to mitigate the effect of estimation errors in $\hat{\boldsymbol{W}}_h$, and to keep a maximum number of entries in $\hat{\boldsymbol{b}}_{I+h}$ unchanged, we propose to compute *sparse adjustments*. An additional regularization will be applied by imposing a smoothness constraint on successive adjustments.

We can compute sparse and smooth adjustments by penalizing the L_1 norm of both the adjustments and the differences in successive adjustments. This is also known as the *sparse fused lasso* (Tibshirani et al., 2005). Finally, we will penalize each entry in the adjustment vector adaptively depending on the magnitude of the `MinT` adjustments.

In other words, we will solve a sparse fused adaptive lasso (SFAL) problem, and compute our adjustments using

$$\hat{\boldsymbol{\theta}}^{\text{SFAL}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^m}{\operatorname{argmin}} \left\{ (\hat{\boldsymbol{z}}_{I+h} - \boldsymbol{K}\boldsymbol{\theta})' \hat{\boldsymbol{W}}_h^{-1} (\hat{\boldsymbol{z}}_{I+h} - \boldsymbol{K}\boldsymbol{\theta}) + \lambda_1 \left\|\boldsymbol{D}_1\boldsymbol{\theta}\right\|_1 + \lambda_2 \left\|\boldsymbol{D}_2\boldsymbol{\theta}\right\|_1 \right\},$$

where $\lambda_1, \lambda_2 \geq 0$ are regularization parameters, and $\boldsymbol{D}_1 \in \mathbb{R}^{(m-1)\times m}$ and $\boldsymbol{D}_2 \in \mathbb{R}^{m\times m}$ are the penalty matrices associated to the fusion and lasso penalty, respectively.

The fusion penalty matrix is $\boldsymbol{D}_1 = \boldsymbol{U}\boldsymbol{D}$ where $\boldsymbol{U} \in \mathbb{R}^{(m-1)\times(m-1)}$ is a diagonal matrix with diagonal entries $\boldsymbol{U}_{jj} = \frac{1}{|\hat{\theta}_{j+1}^{\text{MinT}} - \hat{\theta}_j^{\text{MinT}}|}$ with $j = 1, \ldots, m-1$, and $\boldsymbol{D} \in \mathbb{R}^{(m-1)\times m}$ is the (first) difference matrix given by

$$\boldsymbol{D} = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix}.$$

The lasso penalty matrix is given by $\boldsymbol{D}_2 = \operatorname{diag}\left(\frac{1}{|\hat{\theta}_1^{\text{MinT}}|}, \ldots, \frac{1}{|\hat{\theta}_m^{\text{MinT}}|}\right)$.

The fusion and lasso penalties can also be written as

$$\left\|\boldsymbol{D}_1\boldsymbol{\theta}\right\|_1 = \sum_{j=2}^{m} \frac{1}{|\hat{\theta}_j^{\text{MinT}} - \hat{\theta}_{j-1}^{\text{MinT}}|} |\theta_j - \theta_{j-1}|, \tag{11}$$

$$\left\|\boldsymbol{D}_2\boldsymbol{\theta}\right\|_1 = \frac{1}{|\hat{\theta}_j^{\text{MinT}}|} \sum_{j=1}^{m} |\theta_j|. \tag{12}$$

We can see in (11) that the fusion penalty assigns larger penalties to the successive adjustments that are more similar to each other, which implies the differences are shrunk toward zero faster. If the successive adjustments are significanlty different, a smaller weight is then assigned, which allows a larger change in the differences. A similar idea is used for the lasso penalty in (12). Note that since $\boldsymbol{\theta} = \tilde{\boldsymbol{b}}_{I+h} - \hat{\boldsymbol{b}}_{I+h}$, a shrinkage of the adjustment $\boldsymbol{\theta}$ towards zero is equivalent to a shrinkage of the bottom revised forecasts $\tilde{\boldsymbol{b}}_{I+h}$ towards the bottom base forecasts $\hat{\boldsymbol{b}}_{I+h}$.

When $\lambda_1 = \lambda_2 = 0$, $\hat{\boldsymbol{\theta}}^{\text{SFAL}}$ in (10) and $\hat{\boldsymbol{\theta}}^{\text{MinT}}$ in (4) are equal, and when $\lambda_1 = \infty$ or $\lambda_2 = \infty$, there will be no adjustments as in `BU`, i.e. $\hat{\boldsymbol{\theta}}^{\text{SFAL}} = \mathbf{0} = \hat{\boldsymbol{\theta}}^{\text{BU}}$. The goal of course is the find the right trade-off between these two extremes by properly choosing the values of λ_1 and λ_2 that minimize the forecast errors.

If $\hat{\boldsymbol{W}}_h = \boldsymbol{C}\boldsymbol{C}'$, the optimzation problem in (4) can also be formulated as a *generalized lasso* problem (Tibshirani and Taylor, 2011). The adjustments can then be computed as

$$\hat{\boldsymbol{\theta}}^{\text{SFAL}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^m}{\operatorname{argmin}} \left\{ \left\| \hat{\boldsymbol{l}}_{I+h} - \boldsymbol{X}\boldsymbol{\theta} \right\|_2^2 + \lambda \left\| \boldsymbol{G}\boldsymbol{\theta} \right\|_1 \right\}, \tag{13}$$

where $\hat{\boldsymbol{l}}_{I+h} = \boldsymbol{C}^{-1}\hat{\boldsymbol{z}}_{I+h}$, $\boldsymbol{X} = \boldsymbol{C}^{-1}\boldsymbol{K}$, $\lambda \geq 0$ is a regularization parameter, and $\boldsymbol{G} = \begin{bmatrix} \boldsymbol{D}_1 \\ \gamma \boldsymbol{D}_2 \end{bmatrix}$ with $\gamma \geq 0$ indicating the ratio of the lasso and fusion regularization.

For fixed values of λ and γ, the solution can be computed using a convex optimization solver. Unfortunately, the efficient coordinate descent method often used to solve large-scale lasso problems (Friedman et al., 2007) is not guaranteed to converge in our case since the penalty term is not separable in $\boldsymbol{\theta}$.

We will consider the regularization path algorithm developed in Tibshirani and Taylor (2011), which computes the solution of the generalized lasso for all values of the tuning parameter λ simultaneously. It is based on solving the dual of the generalized lasso, which simplifies the computation of the path (Arnold and Tibshirani, 2016). In other words, for a fixed value of γ, we compute the solutions $\hat{\boldsymbol{\theta}}_{\gamma}^{\text{SFAL}}(\lambda)$ in (13) for $\lambda \in (0, \infty]$. Although it is not the most efficient approach in large scale problems, it is appropriate in the context of temporal hierarchies since the number of observations M in (13) will be small even with a high-frequency time series.

5. Experiments

5.1 Experimental Setup

We consider the smart meter data set used in Ben Taieb et al. (2017) to evaluate multiple forecasting methods for contemporaneous hierarchies. The data set contains half-hourly measurements of electricity consumption gathered from over $14,000$ households from January 2008 to September 2010. We focus on 5000 meters which do not have missing values, with data available between April 20, 2009 and July 31, 2010; hence, each time series has $T = 22,464$ observations. This dataset is particularly suitable for temporal aggregation due to the high-frequency time series and large number of observations.

We focus on one-day ahead demand forecasting, i.e. we generate 48 half-hour forecasts for the next day with a forecast origin at 23:30 ($m = 48$ and $H = 1$). For each half-hourly series, we compute the k-aggregate series for all factors of $m = 48$, from daily to half-hourly observations ($k \in \{48, 24, 16, 12, 8, 6, 4, 3, 2, 1\}$). The length of the k-aggregate series is $T_k = \frac{22,464}{k}$ with respective forecast horizon being $H_k \in \{1, 2, 3, 4, 6, 8, 12, 16, 24, 48\}$. We construct a temporal hierarchy, as proposed in Section 2, and end up with $I = 468$ historical observations where $\boldsymbol{b}_i \in \mathbb{R}^{48}$, $\boldsymbol{y}_i \in \mathbb{R}^{124}$ and $i = 1, \ldots, I$. Finally, we split the data into training, validation and test sets; the first 12 months with 346 days for training, the next month with 30 days for validation and the remaining months with 92 days for

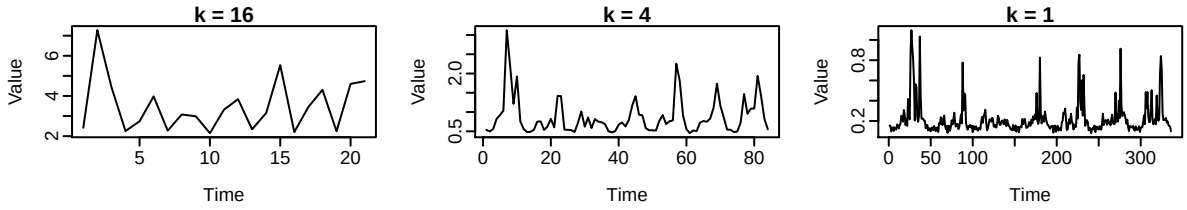

Figure 1: Multiple temporal aggregation for a week of observations.

testing. Figure 1 gives the temporal hierarchy for one series over one week at multiple levels of aggregation.

5.2 Forecasting methods

We will compare our forecasting algorithm `SFAL` with `BU`, `BASE` and `MinT`. The base forecasts are independently generated for each series at each aggregation level. For the lower level series where $k \in \{2, 1\}$, we generate the base forecasts using an exponential smoothing based method, called TBATS (Livera et al., 2011) that allows to capture the within-day and within-week seasonalities in the half-hourly and hourly demand. The parameters of the TBATS model are estimated by maximum likelihood and model selection is performed using AIC. For the upper level series where $k \in \{48, 24, 16, 12, 8, 6, 4, 3\}$, we use the automated forecasting algorithms based on exponential smoothing and ARIMA models, as described in Hyndman and Khandakar (2008). Finally, in order to stabilize the variance and guarantee the non-negativity of the base forecasts, we apply a log transformation.

For both `SFAL` and `MinT`, we estimate the matrix $\boldsymbol{W}_h$ using the shrinkage estimator proposed by Schäfer and Strimmer (2005) with a block-diagonal target. The estimate is recomputed every 10 days in both validation and test sets. The same estimator has been considered in Wickramasuriya et al. (2015) for contemporaneous hierarchies.

The `SFAL` forecasts in (13) depends on two regularization parameters λ and γ that control the smoothness and sparsity of the adjustments. We consider each value of γ in the lasso regularization path, and for a given value of γ, we generate the regularizaion path of the associated generalized lasso problem. We then select the best values of the two parameters by minimizing validation errors.

5.3 Forecast Evaluation

Given the forecasts $\hat{\boldsymbol{y}}_i^{[k]}$ and the actual observations $\boldsymbol{y}_i^{[k]}$ for day i at level k, we compute the mean squared forecast error (MSFE) at level k as

$$\text{MSFE}(k) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left\| \hat{\boldsymbol{y}}_i^{[k]} - \boldsymbol{y}_i^{[k]} \right\|_2^2, \tag{14}$$

where n_{test} is the number of days in the test set. This is equivalent to the MSFE averaged over the M_k forecast horizons and the n_{test} different forecast origins.

The base forecasts at each level of aggregation form a natural benchmark. However, these base forecasts are not necessarily coherent, and do not take into account information

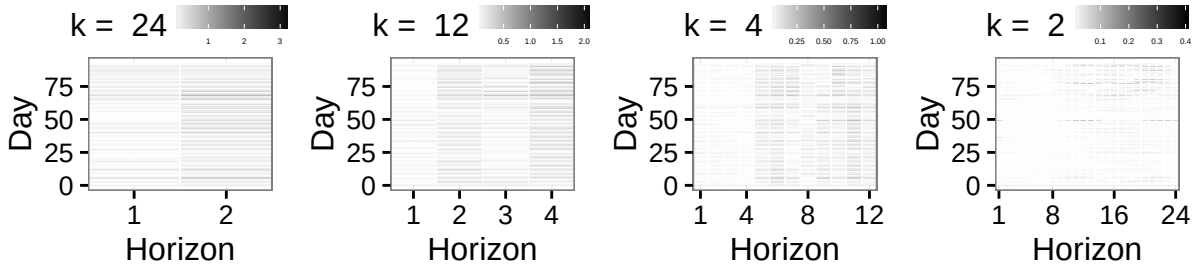

Figure 2: Absolute coherency errors over the forecast horizon for all days in the test set at multiple levels of aggregation.

at other levels. Although the revised forecasts are coherent, an additional desirable property would be that the revised forecasts are at least as accurate as the base forecasts.

In order to quantify the gain/loss of the revised forecasts with respect to the base forecasts, we consider the *Skill Score* (SKS) at level k, which is defined as

$$\text{SKS}(k) = \frac{\text{SCORE}_{\text{BASE}}(k) - \text{SCORE}(k)}{\text{SCORE}_{\text{BASE}}(k)} = 1 - \frac{\text{SCORE}(k)}{\text{SCORE}_{\text{BASE}}(k)}, \tag{15}$$

where $\text{SCORE}(k) = \text{MSFE}(k)$. Finally, the SKS of the 5000 series is averaged to obtain an *Average Skill Score* (ASKS) given by $\text{ASKS}(k) = \frac{1}{5000}\sum_{j=1}^{5000} \text{SKS}_j(k)$ where $\text{SKS}_j(k)$ is the skill score of the jth series at level k.

5.4 Results

Figure 2 shows the coherency errors defined in (1) in absolute value for one series of the smart meter data set. Each panel is associated to one level of aggregation and gives the absolute coherency errors for each day in the test set (row) over the forecast horizon (column). We can see that the magnitude of the coherency errors increases with the aggregation level. This is expected since the magnitude of the forecast errors also increases with the aggregation level. For the considered series, the first horizons have lower coherency errors which can be explained by the lower demand during night hours compared to day hours.

Table 1 gives the average skill score (in %) for the different methods, while the associated bootstrapped standard errors are given in Table 2. Recall that the skill score represents the percentage decrease in MSFE with respect to the base forecasts, i.e. the higher the value the better.

Table 1 shows that the `BU` forecasts are outperformed by the base forecasts at all aggregation levels with a higher decrease in performance for higher levels. Recall that `BU` does not apply any adjustment to the base forecasts before aggregation, and hence it is an extreme case of `SFAL` with $\lambda = \infty$ in (13).

In contrast to `BU`, both `MinT` and `SFAL` improve over the base forecasts for the first k-aggregate series with $k \in \{1, 2, 3, 4\}$. This suggests that the combination of forecasts from multiple aggregation levels helps decrease the forecast errors in the first aggregation levels. However, the magnitude of improvements generally decreases with the level. In particular,

Table 1: Average skill score (in %) at different levels of aggregation.

ASKS	k									
	1	2	3	4	6	8	12	16	24	48
BU	0.00	−1.72	−14.03	−26.54	−37.85	−41.80	−52.50	−43.83	−60.03	−90.8
MinT	5.21	10.46	9.35	1.93	−3.47	−5.71	−8.19	−5.11	−9.27	−13.3
SFAL	8.12	11.51	9.33	1.54	−4.00	−6.18	−8.94	−5.28	−9.38	−13.8

Table 2: Bootstrapped standard errors associated to Table 1.

SE	k									
	1	2	3	4	6	8	12	16	24	48
BU	0.000	0.195	0.963	0.962	1.222	1.293	1.434	1.480	1.80	2.58
MinT	0.295	0.236	0.450	0.305	0.617	0.799	0.918	0.974	1.08	1.14
SFAL	0.261	0.216	0.451	0.307	0.607	0.759	0.892	0.927	1.00	1.07

both methods no longer improve over the base forecasts for $k > 4$; however the decrease in accuracy is significantly smaller than that of `BU`.

The higher errors for the aggregated forecasts compared to the base forecasts is expected since the bottom forecasts often have a higher variance due to the low signal-to-noise ratio of the bottom series. The base forecasts for the aggregated series benefit from lower forecast variance due to the higher signal-to-noise ratio. Nevertheless, compared to the base forecasts, the aggregated forecasts are coherent. This suggests that there is a fundamental tradeoff between forecast accuracy at higher aggregation levels and the coherency of the forecasts.

Finally, if we compare `SFAL` with `MinT`, we can see that `SFAL` has higher skill score than `MinT` for $k \in \{1, 2\}$ and comparable skill score for $k > 3$. This shows that `SFAL` is able to generate forecasts at least as accurate as `MinT` at higher aggregation levels, and at the same provide sparse adjustments with better forecast accuracy at the bottom level.

The first panel of Figure 3 shows the histogram of the number of adjustments in 48 half-hours averaged over the test set for each of the 5000 series. We can see that `SFAL` generates revised forecasts with more than 10 (out of 48) half-hours without adjustments for about 60% of the time series.

Finally, to illustrate the sparsity in the adjustments, the middle and right panels of Figure 3 compare the adjustments applied to the base forecasts by `MinT` (middle panel) and `SFAL` (right panel) for one time series for each day in the test set (row) at each half-hour (column).

6. Conclusion

We proposed a new algorithm to generate multi-period ahead coherent forecasts for multiple temporal aggregation of a given time series. By applying adjustments to possibly incoher-

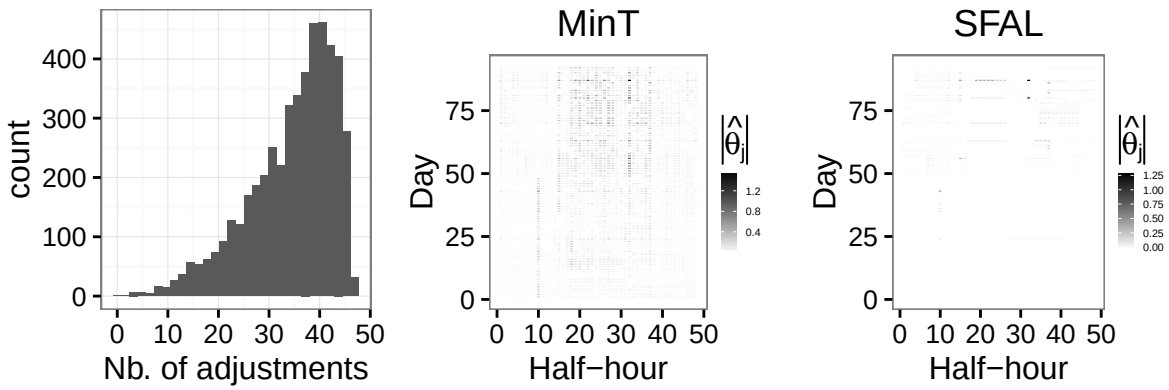

Figure 3: Histogram of the average number of adjustments in 48 half-hours (left panel). Absolute value of adjustments for MinT (middle panel) and SFAL (right panel).

ent base forecasts, the algorithm provides revised forecasts that satisfy the aggregation constraints.

Since these adjustments are applied to consecutive observations, a fusion penalty is added to the objective function in order to obtain smooth adjustments. An additional lasso penalty allows for the estimation of sparse adjustments. These regularised adjustments are computed by solving a generalized lasso problem. Such a regularization allows a reduction in the number of adjustments applied to the base forecasts, and brings more robustness to estimation errors in the covariance of the base forecast errors.

The experiments performed on a large-scale smart meter dataset confirm the effectiveness of the proposed algorithm compared to the state-of-the art methods. In particular, our algorithm produces daily adjustments that are about 20% sparser compared to the benchmark method, and provide even better forecast accuracy in the first aggregation levels.

In addition to electricity demand, many other applications, from renewable energies to tourism demand, will also benefit from more accurate and coherent temporally aggregated forecasts.

References

Taylor B Arnold and Ryan J Tibshirani. Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1):1–27, 2016.

George Athanasopoulos, Rob J Hyndman, Nikolaos Kourentzes, and Fotios Petropoulos. Forecasting with temporal hierarchies. Technical Report 16/15, 2015.

Souhaib Ben Taieb, Jiafan Yu, Barreto Mateus N., and Ram Rajagopal. Regularization in hierarchical time series forecasting with application to electricity smart meter data. In *Thirty-First AAAI Conference on Artificial Intelligence*, 1 March 2017.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, December 2007.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, June 2011.

Rob Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(1):1–22, 2008.

Rob J Hyndman, Roman A Ahmed, George Athanasopoulos, and Han Lin Shang. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579–2589, 1 September 2011.

Rob J Hyndman, Alan J Lee, and Earo Wang. Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis*, 97: 16–32, May 2016.

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.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:32, 14 November 2005.

Andrea Silvestrini and David Veredas. Temporal aggregation of univariate and multivariate time series models: a survey. *Journal of Economic Surveys*, 22(3):458–497, 1 July 2008.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 67(1):91–108, 1 February 2005.

Ryan J Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, June 2011.

Shanika L Wickramasuriya, George Athanasopoulos, and Rob J Hyndman. Forecasting hierarchical and grouped time series through trace minimization. Technical Report 15/15, 2015.