The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.2.229
31(2), 229–239

Temporal hierarchical forecasting with an application to traffic accident counts

Gwanyoung Jun[a] · Byeongchan Seong[a,1]

[a]Department of Applied Statistics, Chung-Ang University

(Received January 16, 2018; Revised February 24, 2018; Accepted February 24, 2018)

Abstract

This paper introduces how to adopt the concept of temporal hierarchies to forecast time series data. Similarly as in hierarchical cross-sectional data, temporal hierarchies can be constructed for any time series data by means of non-overlapping temporal aggregation. Reconciliation forecasts with temporal hierarchies result in more accurate and robust forecasts when compared with the independent base and bottom-up forecasts. As an empirical example, we forecast traffic accident counts with temporal hierarchies and observe that reconciliation forecasts are superior to the base and bottom-up forecasts in terms of forecast accuracy.

Keywords: temporal hierarchies, reconciliation forecast, weighted least square estimator, ARIMA model, exponential smoothing method

1. 서론

국가의 정책이나 기업의 의사결정 등에 사용되는 거시적인 의사결정에 필요한 시계열 자료들은 상대적으로 시간의 주기 단위가 길다(예: 년 또는 분기 등). 이러한 시간 단위가 긴 자료의 예측의 경우 보통 경험이나 전문가의 지식을 바탕으로 예측하는 것이 유리하다. 하지만 내일 판매량과 몇 초 후의 전력 수요량 등 시간 단위가 짧은 자료의 경우 과거 자료를 통한 모형 적합 및 예측이 더욱 유리하다. 일반적으로 시간 단위가 긴 자료의 예측에서는 추세가 잘 반영되지만 계절성이 잘 반영되지 못하는 경우가 많으며, 이와는 반대로 시간 단위가 짧은 자료의 경우 계절성은 잘 반영되지만 추세 반영이 부족한 경우가 많다. 본 논문에서 소개할 시간적 계층을 이용한 예측(forecasting with temporal hierarchies)은 시계열 자료가 가지는 다양한 주기의 시간 계층 구조를 활용하여 자료를 예측 및 조정함으로써 예측 시간 단위의 장단에 관계없이 추세와 계절성을 모두 반영할 수 있는 장점을 제공한다. 즉, 시간 주기의 장단에 따라 서로 상이한 예측값들을 하나로 통합 또는 조합하는 효과를 가지고 있다.

시간적 계층을 이용한 예측은 비교적 최근에 관심을 받게 된 주제로 다양한 분야에서 사용되고 있다. Athanasopoulos 등 (2017)은 영국의 주별(weekly) 재해 및 긴급 구조 수요(demand of accident and emergency) 자료를 예측하기 위하여 시간적 계층 구조를 이용하였다. 주별 자료는 6개의 시간적 계층으로 구조화되어 수요 예측에 활용되었다.

This research was supported by the Chung-Ang University Research Scholarship Grants in 2016 and it is a revision of the first author's master's thesis.

[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: bcseong@cau.ac.kr

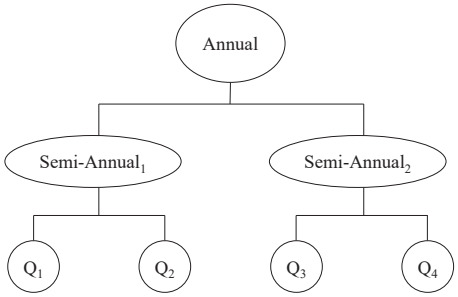

Figure 2.1. Temporal hierarchical structure for quarterly time series.

시간적 계층을 이용한 예측 방법은 주로 상향식(bottom-up) 예측과 조정 예측(reconciliation forecast) 방법이 널리 사용된다. 상향식 예측은 최하위 계층의 시계열 자료의 예측값을 이용해 이를 중복되지 않게 더하면서 상위 단계의 예측값을 생성하는 방법이다. Athanasopoulos 등 (2009)이 제안한 조정 예측은 선형 회귀모형을 기반으로 최소제곱 추정량을 사용하여 시간적 계층의 예측값을 추정한다. 이러한 시간적 계층을 사용한 예측 방법을 사용하면 항상 하위 계층의 예측값의 합은 상위 계층의 예측값과 일치하게 된다 (Lee와 Seong, 2017).

본 논문에서는 시간적 계층을 이용하여 교통사고 발생건수를 예측하고, 기존의 방식인 기저 예측(base forecast) 및 상향식 예측과 비교하였다. 국내의 주요 선행 연구들을 살펴보면, 교통사고 관련 시계열 자료를 예측할 때 단변량 모형이 주로 사용되었다. Kim과 Lee (2014)는 충청도 주요 도시의 교통사고 발생건수를 autoregressive integrated moving average (ARIMA) 모형을 사용해 예측하고 적정성을 검토하였으며, Han (2007)은 도로 종류별 교통사고 발생건수를 ARIMA 모형을 사용해 예측하였다. 즉, 모든 선행 연구들은 정해진 한가지의 시간 주기를 통한 예측에만 국한되어 있다.

본 논문에서는 시간적 계층을 이용한 예측 방법을 소개하고, 실증 분석을 통하여 그 성능을 기저 예측 및 상향식 방법과 비교한다. 논문은 총 4장으로 구성되어 있으며, 2장에서는 시간적 계층을 정의하고 이를 활용한 예측 방법을 설명하고, 3장에서는 이를 이용하여 국내 교통사고 발생건수 시계열 자료를 월별부터 연별까지 다양한 시간적 계층으로 예측하고 그 성능을 비교하였고, 4장에서는 결론을 맺는다.

2. 시간적 계층을 이용한 예측

2.1. 시간적 계층

시간적 계층은 시계열 자료의 부분합을 이용하여 만들며 여러 단계의 계층으로 이루어진다. 최하위 단계는 통합되지 않은 원시계열 자료를 나타내며 이러한 하위 계층의 시계열 자료들의 부분합으로 상위 계층이 만들어진다. 예를 들어, Figure 2.1은 분기별 시계열 자료의 시간적 계층 구조를 표현한 것이며, 중복되지 않게 2개씩 분기별 시계열 자료를 더하여 반년별 계층을 구성하고, 반년별 계층의 원소들의 부분합으로 연별 계층의 시계열 원소들을 생성하고 있다. 원시계열 자료가 $\{y_t : t = 1, 2, \ldots, T\}$라고 할 때, 시계열 계층의 각 원소들은 다음과 같이 표현할 수 있다.

$$y_j^{[k]} = \sum_{t=t^*+(j-1)k}^{t^*+jk-1} y_t, \quad j = 1, \ldots, \lfloor T/k \rfloor \ , \ t^* = T - \lfloor T/m \rfloor m + 1, \tag{2.1}$$

여기서 k는 시간적 계층의 수준(level), m은 원시계열의 계절 주기(seasonal period 또는 highest avail-

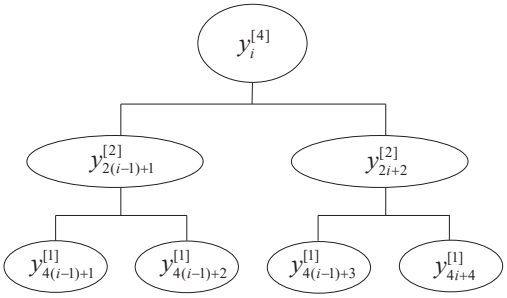

Figure 2.2. Temporal hierarchical structure for quarterly time series using the common index i.

able sampling frequency), $\lfloor\ \rfloor$는 가우스 기호를 나타낸다. 시간적 계층 수준 k는 m의 약수로 이루어져 있다. 즉, p를 시간적 계층의 전체 개수라고 할 때 k는 집합 $\{k_p, \ldots, k_2, k_1\}$의 원소가 된다 (단, $k_p = m$, $k_1 = 1$). 예를 들어, 원시계열이 월별 자료인 경우 $m = 12$, k는 $k \in \{12, 6, 4, 3, 2, 1\}$이 된다.

총 $\lfloor T/m \rfloor$개의 계절에 대하여 i번째 계절에서 k-수준 시간적 계층의 시계열을 벡터 $\mathbf{Y}_i^{[k]}$로 집약하면 다음과 같은 표현이 가능하다.

$$\mathbf{Y}_i^{[k]} = \left(y_{M_k(i-1)+1}^{[k]}, y_{M_k(i-1)+2}^{[k]}, \ldots, y_{M_k i}^{[k]}\right)', \quad i = 1, \ldots, \lfloor T/m \rfloor,\ M_k = \frac{m}{k}. \tag{2.2}$$

또한, 총 p개의 시간적 계층에 해당하는 $\mathbf{Y}_i^{[k]}$들을 모두 쌓아서 다음과 같은 벡터 표현이 가능하다.

$$\mathbf{Y}_i = \left(y_i^{[m]}, \ldots, \mathbf{Y}_i^{[k_3]\prime}, \mathbf{Y}_i^{[k_2]\prime}, \mathbf{Y}_i^{[1]\prime}\right)', \quad i = 1, \ldots, \lfloor T/m \rfloor, \tag{2.3}$$

여기서 $y_i^{[m]} = \mathbf{Y}_i^{[k_p]}$임에 주의하여라. 앞에서 설명한 모든 계층의 벡터가 포함된 $\mathbf{Y}_i$는 Hyndman 등 (2011)이 사용한 합계 행렬(summing matrix)과 최하위 계층의 시계열 벡터 $\mathbf{Y}_i^{[1]}$를 사용하여 아래와 같이 표현할 수 있다. 단, 합계 행렬은 $(K \times m)$차원을 가지는 행렬로서 K는 집합 $\{k_p, \ldots, k_2, k_1\}$의 원소들의 합이다.

$$\mathbf{Y}_i = S\mathbf{Y}_i^{[1]}. \tag{2.4}$$

예를 들어 Figure 2.2의 시간적 계층은 합계 행렬을 사용하여 다음과 같이 표현할 수 있다.

$$\mathbf{Y}_i = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{4(i-1)+1}^{[1]} \\ y_{4(i-1)+2}^{[1]} \\ y_{4(i-1)+3}^{[1]} \\ y_{4i}^{[1]} \end{bmatrix} \quad \left(= S\mathbf{Y}_i^{[1]}\right).$$

2.2. 시간적 계층을 이용한 예측 방법

기저 모형(base model)을 이용한 시간적 계층의 예측 방법은, 원시계열로 만든 p개의 시간적 계층 각각에 대하여 개별적인 예측값(기저 예측)을 사용하여 각 계층을 예측하는 것을 의미한다 (Hyndman과

Athanasopoulos, 2014). 여기서, 개별적인 예측값은 임의의 예측 방법을 사용할 수 있으나 보통은 제어의 용이성 및 자동화 가능성을 고려하여 단변량 시계열 예측 방법인 ARIMA 모형 또는 지수평활법(exponential smoothing method)을 사용한다.

원시계열을 기준으로 기저 예측값을 위한 최대 예측 시차(forecast horizon)를 h^*라고 할 때 최상위 계층에서의 예측 시차 h는 $h = 1, \ldots, \lfloor h^*/m \rfloor$로 표현할 수 있으며, 각 k-수준의 시간적 계층에서는 $\lfloor T/k \rfloor$개의 시계열을 토대로 예측 시차가 $M_k h$인 기저 예측값을 구할 수 있다. 이와 같은 모든 계층에서의 예측 시차 h의 기저 예측값을 식 (2.3)과 같이 한 개의 벡터로 모아서 표현하면 다음과 같다.

$$\hat{\mathbf{Y}}_h = \left(\hat{y}_h^{[m]}, \ldots, \hat{\mathbf{Y}}_h^{[k_3]\prime}, \hat{\mathbf{Y}}_h^{[k_2]\prime}, \hat{\mathbf{Y}}_h^{[k_1]\prime}\right)', \tag{2.5}$$

여기서 $\hat{\mathbf{Y}}_h^{[k]} = (\hat{y}_{M_k(h-1)+2}^{[k]}, \hat{y}_{M_k(h-1)+1}^{[k]}, \ldots, \hat{y}_{M_k h}^{[k]})'$는 M_k차원의 벡터이며 $\hat{\mathbf{Y}}_h$는 $\sum_{\ell=1}^{p} k_\ell$차원의 벡터이다.

기저 예측값 $\hat{\mathbf{Y}}_h$는 합계 행렬 S를 사용하여 다음과 같이 표현할 수 있다.

$$\hat{\mathbf{Y}}_h = S\beta(h) + \varepsilon_h. \tag{2.6}$$

즉, $\beta(h) = E[\mathbf{Y}_{\lfloor T/m \rfloor + h}^{[1]} \mid y_1, \ldots, y_T]$는 최하위 계층의 예측값에 대한 미지의 평균이며, ε_h은 조정 오차(reconciliation error)로서 기저 예측값과 이것의 기대값과의 차이를 나타낸다고 볼 수 있다. ε_h는 평균이 0이고 분산-공분산 행렬 Σ를 가진다고 가정하며, 식 (2.6)은 종단적 조정 회귀모형(temporal reconciliation regression model)이라고 부른다. 이러한 접근 방법은 Hyndman 등 (2011)에서 제안된 횡단적 계층 조정 회귀모형(cross-sectional hierarchical reconciliation regression model)과 유사한 구조를 가지고 있다. 식 (2.6)은 횡단적 계층 모형의 구조를 종단적 구조에 확장한 것으로 볼 수 있다.

공분산 행렬 Σ를 알고 있을 경우, $\beta(h)$의 일반화 최소제곱 추정량(generalized least squares estimator)을 사용하여 기저 예측값 $\hat{\mathbf{Y}}_h$의 적합값 $\tilde{\mathbf{Y}}_h$을 다음과 같이 구할 수 있다.

$$\tilde{\mathbf{Y}}_h = S\hat{\beta}(h) = S\big(S'\Sigma^{-1}S\big)^{-1}S'\Sigma^{-1}\hat{\mathbf{Y}}_h = SP\hat{\mathbf{Y}}_h. \tag{2.7}$$

단, $P = (S'\Sigma^{-1}S)^{-1}S'\Sigma^{-1}$이다. 위에서 구한 적합값 $\tilde{\mathbf{Y}}_h$를 조정 예측값이라고 부른다. 기저 예측값 $\hat{\mathbf{Y}}_h$가 최소제곱법에 의하여 조정되었다는(reconciled) 관점으로 지어진 이름이다. 그러나, 분산-공분산 행렬 Σ은 일반적으로 미지이므로 추정하여야 하지만 식별할 수 없다(not identifiable) (Wickramasuriya 등, 2015). 대안으로서, 각 시간적 계층에서의 예측이 서로 독립이라는 가정하에서, Athanasopoulos 등 (2009)은 Σ의 추정량으로서 $W = \sigma^2 I$와 같이 대각화 공분산 행렬 구조를 사용하였으며, Athanasopoulos 등 (2017)은 다음 절에서 소개할 가중 최소제곱법(weighted least squares estimator)을 제안하였다. Σ의 추정량을 W라고 할 경우, 조정 예측값은 다음과 같이 표현될 수 있다.

$$\tilde{\mathbf{Y}}_h = S\big(S'W^{-1}S\big)^{-1}S'W^{-1}\hat{\mathbf{Y}}_h. \tag{2.8}$$

2.3. 분산-공분산 행렬 Σ의 추정

본 절에서는 W의 계산을 위하여 Athanasopoulos 등 (2017)이 제안한 방법 중 두 가지 방법을 소개한다. 두 방법 모두 각 시계열 계층에서의 기저 예측이 독립이라고 가정하고 추가적인 가정을 한다.

2.3.1. 시간적 계층별 등분산 가정 각 시계열 계층에서의 기저 예측이 독립이라고 가정한다면 W를 대각행렬로 계산할 수 있다. 이러한 방법은 Σ를 일반적으로 추정하는 것보다 훨씬 적은 개수의 모수를

추정하게 된다. 또한 일반적인 시계열 모형처럼 동일한 시간적 계층 내에서 등분산을 가진다는 가정을 한다면 추정할 모수의 개수를 더 줄일 수 있다. 시간적 계층별 등분산(series variance scaling) 가정 하에서는 동일한 시간적 계층 내에서 W의 원소들은 같은 값을 가진다. 예를 들어 Figure 2.1와 같은 구조에서 W는 아래와 같은 형태를 가진다.

$$W_V = \text{diag}\Big(\hat{\sigma}^{[4]}, \hat{\sigma}^{[2]}, \hat{\sigma}^{[2]}, \hat{\sigma}^{[1]}, \hat{\sigma}^{[1]}, \hat{\sigma}^{[1]}, \hat{\sigma}^{[1]}\Big)^2. \tag{2.9}$$

단, $\hat{\sigma}^{[i]}$는 i-계층에서의 조정 오차 분산의 추정값을 나타내고, diag( )^{2}은 제곱한 각 원소들을 대각 원소로 하는 행렬을 의미한다.

2.3.2. 구조적 등분산 가정 시간적 계층별 등분산 가정은 분산-공분산 행렬의 추정을 위하여 p개의 모수를 추정한다. 그러나, 추가적으로 추정할 모수의 개수를 줄일 수 있다. 기본적으로 시간적 계층의 각 수준별 시계열은 최하위 계층의 시계열의 합을 통하여 만들어지는 구조이다. 따라서 최하위 계층의 시계열 자료의 조정 오차의 추정 분산을 $\hat{\sigma}^2$이라고 하면, 각 수준별 예측이 서로 독립적이라는 가정 하에서 Σ의 추정량은 다음과 같이 계산될 수 있다.

$$W_S = \hat{\sigma}^2 \times \text{diag}(S1). \tag{2.10}$$

단, 1은 $\mathbf{Y}_i^{[1]}$의 차원과 같은 차원을 가지는 단위 벡터(unit vector)이다. 구조적 등분산(structural scaling) 가정을 사용하면 여러 장점이 존재한다. 첫째, Σ의 추정이 최하위 시계열의 계절주기인 m에만 의존하기 때문에 시계열 자료 및 예측 모형에 의존하지 않는다. 둘째, 추정의 과정이 단순하므로 비계량적 예측을 포함하여 모든 예측 방법에 사용할 수 있다. 예를 들어 Figure 2.1을 통해 계산한 W_S는 아래와 같다.

$$W_S = \hat{\sigma}^2 \times \text{diag}(4, 2, 2, 1, 1, 1, 1). \tag{2.11}$$

3. 실증 분석

이 장에서는 실제 시계열 자료를 사용하여 시간적 계층을 구성하고 각 계층을 예측한다. 기저 모형 및 예측을 위하여 자동적으로 ARIMA 모형을 추정해 주는 R의 forecast 패키지 (Hyndman, 2017)의 함수인 auto.arima와 자동적으로 지수평활법을 추정해 주는 ets 함수를 사용하였다. 최적 모형을 찾기 위한 기준으로 Akaike information criterion (AIC) 정보량을 사용하였으며, 최적 모형으로 결정된 기저 모형을 사용하여 다음과 같은 4가지 예측 방법의 각 계층별 예측력을 비교하였다.

(1) 기저 모형을 이용한 예측 방법(BASE)

(2) 최하위 계층의 부분합을 이용하여 예측하는 상향식 방법(BU)

(3) 시간적 계층 예측 방법에서 계층별 등분산을 가정한 경우(W_V)

(4) 시간적 계층 예측 방법에서 구조적 등분산을 가정한 경우(W_S)

시간적 계층 예측은 R의 thief 패키지를 사용하였다 (Hyndman과 Kourentzes, 2016).

3.1. 자료 설명: 국내 교통사고 발생건수

실증 분석에 사용한 자료는, 국내 교통사고 발생건수이며, 1992년 1월부터 2016년 12월까지의 월별 자료이다. 총 시계열 자료의 길이는 300개이며, 모형 적합을 위한 훈련 자료(training set)의 기간은

Table 3.1. Temporal hierarchy of the traffic accident counts

Aggregation level	Length of time-series
Monthly	300
Bi-monthly	150
Quarterly	100
Four-monthly	75
Semi-annual	50
Annual	25

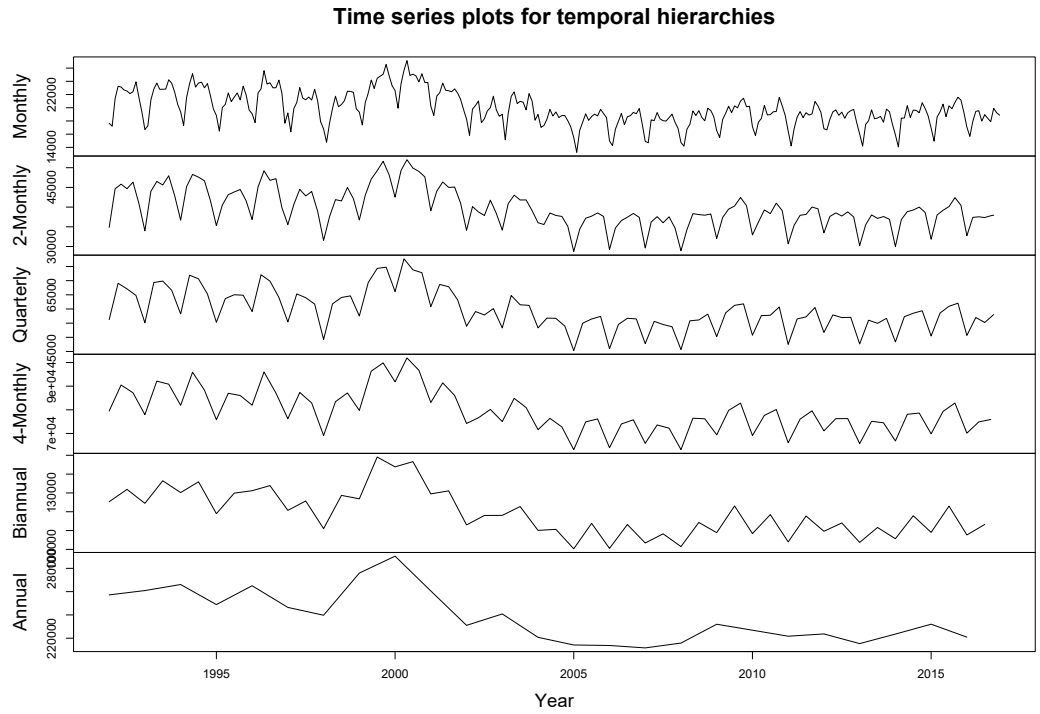

Figure 3.1. Plot of series at each level of temporal hierarchy.

1992년 1월부터 2014년 12월이다. 예측 성능 평가를 위한 검증 자료(test set)의 기간은 2015년 1월부터 2016년 12월까지로 2년간의 자료를 사용하였다. 자료는 교통사고분석시스템(Traffic Accident Analysis System)을 이용하여 얻을 수 있다 (http://taas.koroad.or.kr).

이 자료가 가지고 있는 시간적 계층은 Table 3.1과 같다. 최상위(annual) 수준의 시계열 자료의 길이는 25개로서 연별 자료이고 최하위(monthly) 수준은 길이가 300개인 월별 자료이다. Figure 3.1은 교통사고 발생건수 자료에 대한 시간적 계층 시계열 그림이다. 'Anuual' 그림은 최상위 시간적 계층으로서 연별 교통사고 발생건수를 의미하고 'Biannual' 그림은 반년별 교통사고 발생건수, 'Monthly' 그림은 월별 국내 교통사고 발생 건수를 나타낸다. 하위 단계일수록 계절성의 형태가 분명히 나타나고 있으나 상위 단계로 갈수록 계절성이 사라지는 것을 볼 수 있다. 최상위 단계에서는 계절성을 찾아볼 수 없고 점진적으로 감소하는 대략적인 추세를 확인할 수 있다.

Table 3.2. MAPE and MASE for test set forecasts of the temporal hierarchical time-series methods using ARIMA model applied to the domestic traffic accident counts

		Hierarchical level						
		Monthly	Bi-monthly	Quarterly	Four-monthly	Semi-annual	Annual	Average
MAPE	BASE	6.745	5.796	**2.750**	3.391	**1.911**	**3.148**	3.957
	BU	6.745	6.524	6.280	6.514	6.230	6.094	6.398
	W_v	4.409	4.163	4.211	4.136	4.200	3.686	4.134
	W_s	**3.627**	**3.281**	3.284	**3.079**	3.283	3.149	**3.284**
MASE	BASE	0.957	0.879	**0.449**	0.544	**0.328**	0.584	0.624
	BU	0.957	0.981	0.977	1.022	1.000	1.120	1.009
	W_v	0.636	0.637	0.667	0.659	0.685	0.676	0.660
	W_s	**0.531**	**0.512**	0.528	**0.497**	0.543	**0.582**	**0.532**

MAPE = mean absolute percentage error; MASE = mean absolute scaled error; ARIMA = autoregressive integrated moving average.

Table 3.3. MAPE and MASE for test set forecasts of the temporal hierarchical time-series methods using exponential smoothing method applied to the domestic traffic accident counts

		Hierarchical level						
		Monthly	Bi-monthly	Quarterly	Four-monthly	Semi-annual	Annual	Average
MAPE	BASE	3.971	3.473	**2.964**	**2.771**	2.378	**2.413**	2.995
	BU	3.971	3.618	3.315	3.340	**3.364**	2.509	3.353
	W_V	3.570	**3.173**	3.050	2.915	2.399	2.473	**2.930**
	W_S	**3.566**	3.205	2.994	3.031	2.386	2.458	2.940
MASE	BASE	0.575	0.534	**0.477**	**0.450**	**0.403**	**0.460**	0.483
	BU	0.575	0.556	0.537	0.541	0.551	**0.460**	0.537
	W_V	**0.518**	**0.491**	0.492	0.471	0.403	0.460	**0.473**
	W_S	0.519	0.495	0.482	0.489	0.403	0.460	0.475

MAPE = mean absolute percentage error; MASE = mean absolute scaled error.

3.2. 예측 비교

본 논문에서는 시계열 예측값에 대한 정확성 비교를 위하여 mean absolute percentage error (MAPE)와 mean absolute scaled error (MASE)를 이용하였다. MAPE는 예측값의 퍼센트 오차를 사용하고 MASE는 예측값의 오차와 (계절) 단위근 모형에 의한 예측(seasonal naive forecast)의 오차 간의 비율을 사용하기 때문에 단위나 크기에 관계없이 동일한 기준으로 비교 가능한 장점이 있다. 원시계열 y_j 및 예측값 $\hat{y}_j$에 대하여 미래 예측 시점 h까지의 MAPE와 MASE는 다음과 같이 정의된다.

$$\text{MAPE} = \frac{100}{h}\sum_{j=1}^{h}\left|\frac{y_j - \hat{y}_j}{y_j}\right|, \qquad \text{MASE} = \frac{1}{h}\sum_{j=1}^{h}\frac{|y_j - \hat{y}_j|}{Q}. \tag{3.1}$$

단, $Q = \sum_{t=1}^{T}|y_t - y_{t-m}|/(T-m)$이고 m은 계절 주기를 나타낸다.

Tables 3.2과 3.3은 예측 방법간(BASE, BU, W_V, W_S) 예측력을 비교한 결과이다. Table 3.2는 기저 모형으로 ARIMA 모형을 적합한 경우이고, Table 3.3는 지수평활법으로 적합한 경우이다. 표에서 가장 작은 MAPE 또는 MASE 값을 가지는 예측 방법은 진한 밑줄로 표시하였다. 또한 마지막 열은 모든 계층의 MAPE 또는 MASE 값들의 평균이다. Figures 3.2와 3.3은 각각 Tables 3.2와 3.3을 그래프로 표현한 것이다. Tables 3.4와 3.5는 기저 모형에서 사용된 구체적인 ARIMA 및 exponential

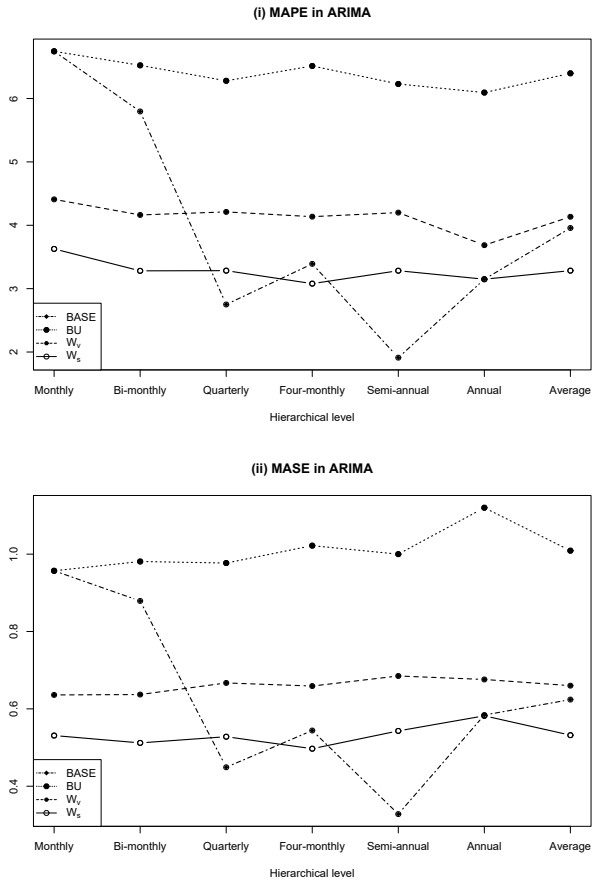

Figure 3.2. Plots of MAPE and MASE for temporal hierarchies using ARIMA model applied to the domestic traffic accident counts. MAPE = mean absolute percentage error; MASE = mean absolute scaled error; ARIMA = autoregressive integrated moving average.

smoothing method (ETS) 모형의 형태를 각각 나타낸다. 두 모형들 모두에서 공통적으로 최상위 계층을 제외한 모든 계층에서 계절 모형을 적합하고 있음을 알 수 있다.

표와 그림을 통하여 기저 모형으로 ARIMA 모형을 사용했을 때, 각 계층별 MAPE와 MASE의 예측 방법간 순위 및 계층별 추세가 비슷한 것을 알 수 있고, 상향식 방법(BU)은 모든 계층에서 다른 예측 방법에 비해서 예측력이 떨어짐을 확인할 수 있다. 분기별, 반년별 계층의 경우 기저 예측(BASE)의 성능이 조정 예측(W_V 또는 W_S)보다 성능이 우수함을 확인할 수 있다. 그러나, MAPE와 MASE 값들의 평균을 살펴보면 W_S(구조적 등분산을 가정한 경우) 방법이 가장 좋은 성능을 나타내고 있으며, Figure 3.2를 통하여 볼 때 조정 예측을 사용하는 경우 MAPE와 MASE의 계층별 변동성(분산)이 기저 예측 방법보다 더 작음을 확인할 수 있다.

기저 모형으로 지수평활법을 사용했을 때에도, 각 계층별 MAPE와 MASE의 추세가 비슷한 것을 알 수 있고, 상향식 방법이 가장 성능이 떨어지는 방법임을 확인할 수 있다. 하지만 총 계층의 MAPE와 MASE의 평균은 W_V(계층별 등분산을 가정한 경우) 방법이 가장 좋은 성능을 나타냈고, 분기별 계층부

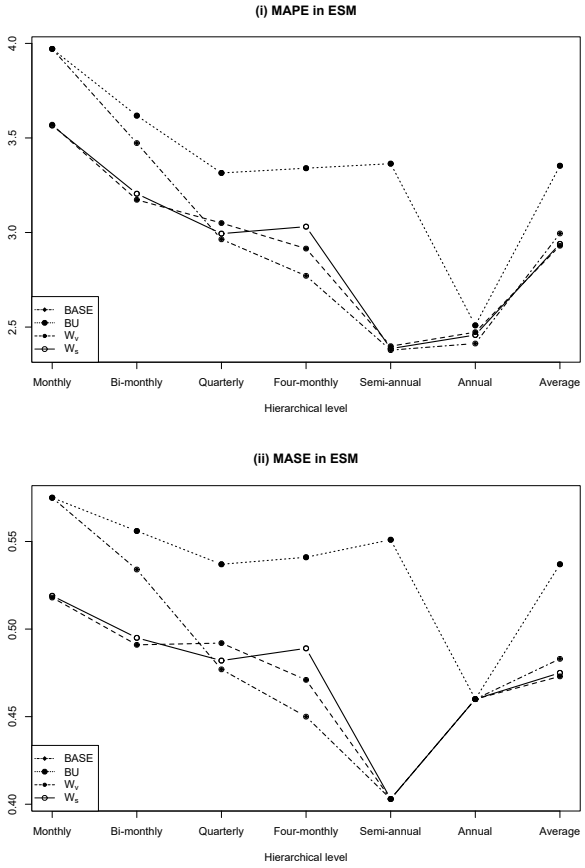

Figure 3.3. Plots of MAPE and MASE for temporal hierarchies using ETS model applied to the domestic traffic accident counts. MAPE = mean absolute percentage error; MASE = mean absolute scaled error; ETS = exponential smoothing method.

Table 3.4. ARIMA models selected as a BASE model in each hierarchical level

Hierarchical level	BASE model
Monthly	$\text{ARIMA}(2,1,0)\times(1,0,0)_{s=12}$
Bi-monthly	$\text{ARIMA}(0,1,0)\times(1,0,0)_{s=6}$
Quarterly	$\text{ARIMA}(2,0,0)\times(0,1,1)_{s=4}$
Four-monthly	$\text{ARIMA}(0,0,2)\times(0,1,0)_{s=3}$
Semi-annual	$\text{ARIMA}(1,0,0)\times(2,1,0)_{s=2}$
Annual	$\text{ARIMA}(1,0,0)$

ARIMA = autoregressive integrated moving average.

터 연별 계층까지 기저 예측이 조정 예측보다 좋은 예측력을 보임을 확인할 수 있었다. ARIMA 모형이 기저 모형일 때와 유사하게 기저 예측은 계층간 변동성이 가장 큰 형태로 나타났다.

전반적으로 국내 교통사고 발생건수는 지수평활법을 이용할 때 전반적으로 우수한 예측력을 보였다. 각 계층별 예측 성능에서 기저 예측이 조정 예측보다 더 좋은 성능을 보일 때도 있으나 계층의 예측 성능의 변동성에 있어서 조정 예측이 더 강건한 예측 방법임을 확인할 수 있었다.

Table 3.5. ETS models selected as a BASE model in each hierarchical level

Hierarchical level	BASE model
Monthly	ETS(A, N, A)
Bi-monthly	ETS(M, N, M)
Quarterly	ETS(M, N, M)
Four-monthly	ETS(M, N, M)
Semi-annual	ETS(M, N, A)
Annual	ETS(M, N, N)

The three characters in parentheses identify method using the framework terminology of Hyndman *et al.* (2002). The first letter denotes the error type (A or M); the second letter denotes the trend type (N, A, or M); and the third letter denotes the season type (N, A, or M). In all cases, N = none, A = additive, and M = multiplicative. ETS = exponential smoothing method.

4. 결론

본 논문에서는 단변량 시계열 자료를 이용하여 시간적 계층을 구성하고 이를 이용하여 각 계층의 예측 성능을 높이는 방법을 소개하고 있다. 실증 분석으로 국내 교통사고 발생건수 월별 자료를 이용하여 다양한 주기의 시간적 계층을 구성한 후 기저 예측으로 ARIMA 모형, 지수평활법을 사용하여 시간적 계층을 이용한 조정 예측 방법의 특징과 장점을 살펴보았다. 그러나 기저 예측으로 동일한 모형만을 사용했기 때문에 향후 각 계층별 예측을 최적화시키는 다양한 모형을 사용하는 연구가 필요하다. 또한 조정 오차의 분산-공분산 행렬을 추정함에 있어 더 정확하고 다양한 추정량을 사용하여 예측값에 계층의 다양성을 반영할 수 있도록 하는 연구가 필요하다고 하겠다.

References

Athanasopoulos, G., Ahmed, R. A., and Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism, *International Journal of Forecasting*, **25**, 146–166.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., and Petropoulos, F. (2017). Forecasting with temporal hierarchies, *European Journal of Operational Research*, **262**, 60–74.

Han, S. J. (2007). Road accident characteristics in metropolitan cities and provinces, *Journal of Environmental Studies*, **46**, 211–220.

Hyndman, R. J. (2017). Forecast: forecasting functions for time series and linear models. R package version 8.2.

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

Hyndman, R. J. and Athanasopoulos, G. (2014). *Forecasting Principles and Practice*, OText, Heathmont.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., and Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting*, **18**, 439–454.

Hyndman, R. J. and Kourentzes, N. (2016). Thief: temporal hierarchical forecasting. R package version 0.2.

Kim, Y. S. and Lee, M. J. (2014). The analysis of predicting traffic accident using ARIMA model. In *Proceeding of the Korean Society of Civil Engineers Autumn Conference*, 705–706.

Lee, J. and Seong, B. (2017). Hierarchical time series forecasting with an application to traffic accident counts, *The Korean Journal of Applied Statistics*, **30**, 181–193.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2015). Forecasting hierarchgical and grouped time series through trace minimization (technical report), Monash University, Melbourne.

시간적 계층을 이용한 교통사고 발생건수 예측

전관영[a] · 성병찬[a,1]

[a]중앙대학교 응용통계학과

(2018년 1월 16일 접수, 2018년 2월 24일 수정, 2018년 2월 24일 채택)

요 약

본 논문에서는 시간적 계층 개념을 활용하여 시계열 자료를 예측하는 방법을 소개한다. 횡단적 계층 자료 분석에서와 유사한 방법으로 중복되지 않는 시간적 계층을 시계열 자료에 구조화할 수 있다. 이러한 시간적 계층을 활용하여 조정된 예측은 기존의 계층별 독립적 기저 예측 및 상향식 예측보다 더 정확하고 강건한 예측값을 생성한다. 실증분석으로서 국내 교통사고 발생건수를 시간적 계층 개념을 활용하여 예측한다. 분석 결과, 조정 예측이 기존의 다른 예측보다 예측 성능면에서 더 우수함을 확인할 수 있다.

주요용어: 시간적 계층, 조정 예측, 가중 최소제곱 추정, ARIMA 모형, 지수평활법

이 논문은 2016년도 중앙대학교 연구 장학기금 지원에 의한 것임. 또한, 제1저자 전관영의 석사학위논문을 수정하여 작성한 것임.

[1]교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 경영경제대학 응용통계학과.
E-mail: bcseong@cau.ac.kr