The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.6.851
32(6), 851–865

Multiple aggregation prediction algorithm applied to traffic accident counts

Doorham Bae[a] · Byeongchan Seong[a,1]

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

(Received August 6, 2019; Revised September 20, 2019; Accepted October 4, 2019)

Abstract

Discovering various features from one time series is complicated. In this paper, we introduce a multi aggregation prediction algorithm (MAPA) that uses the concepts of temporal aggregation and combining forecasts to find multiple patterns from one time series and increase forecasting accuracy. Temporal aggregation produces multiple time series and each series has separate properties. We use exponential smoothing methods in the next step to extract various features of time series components in order to forecast time series components for each series. In the final step, we blend predictions of the same kind of components and forecast the target series by the summation of blended predictions. As an empirical example, we forecast traffic accident counts using MAPA and observe that MAPA performance is superior to conventional methods.

Keywords: temporal aggregation, combination, multiple aggregation prediction algorithm, time series components, exponential smoothing method

1. 서론

현대사회는 빠르게 변화하는 특징을 가지고 있다. 또한, 모든 분야에서 이 빠른 변화에 발맞추어 대응하지 못하면 뒤처지게 되는 구조를 가지고 있다. 이와 같은 빠르게 변화하는 사회에 알맞게 대응하기 위해서는 미래의 변화를 정확하게 예측하는 것이 중요하다. 하지만, 예측을 필요로 하는 다양한 분야들은 여러 가지 요소들에 의해 상호 복합적인 영향을 받으며 이러한 특징은 정확한 예측에 장애물로 작용한다.

이렇듯 복잡한 현상에 대한 예측의 중요성이 커지고 있기 때문에, 예측의 정확성을 높이기 위하여 다양한 방법이 시도되고 있다. 그 중 한 가지는 하나의 자료에서 얻을 수 있는 정보량을 증가시키는 방법이다. 대부분의 시계열 분석 방법은 일변량 자료를 다루며 그 자료의 패턴을 분석하여 미래 시점을 예측한다. 하지만 일변량 자료만을 그대로 사용하는 경우 정보량이 부족할 수 있다. 이 문제를 해결하기 위해 하나의 자료에서 얻을 수 있는 정보를 증가시킬 필요가 있다. 정보량을 증가시키는 방법 중 하나는 시간적 결합(temporal aggregation)을 이용하는 방식이다. 시간적 결합을 이용하여 시계열 자료를 분석 및 예측하는 방법은 꾸준히 연구되는 분야로 Rossana와 Seater (1995)은 미국의 여러 경제 지표를 분석하였고, Rostami-Tabar 등 (2013)은 유럽 식료품 매장의 판매량을 분석하는 등 다양한 분야에서 이용

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2018.
[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: bcseong@cau.ac.kr

되고 있다. 정확한 예측을 위해 사용되는 다른 한 가지 방법은 여러 가지 예측값을 조합하는 방식이다. Trabelsi와 Hillmer (1989)은 다른 주기에서의 autoregressive integrated moving average (ARIMA) 모형을 조합하는 예측 방법을 제시하였고, Casals 등 (2009)은 다른 주기에서의 예측값을 조합하는 방식이 예측 정확도를 향상시킨다는 사실을 이론적으로 증명하였다.

단변량 시계열 분석이 갖는 정보량 부족이라는 한계를 극복하는 방법은 앞서 소개한 시간적 결합의 방식과 시계열 자료의 계층을 활용하는 방식으로 구분할 수 있다. Athanasopoulos 등 (2009)은 계층적 시계열 자료의 예측 방법을 제시하였으며, Hyndman 등 (2011)은 계층형 시계열 자료 예측의 최적 조합 방법을 제시하였다. 또한, Athanasopoulos 등 (2017)은 시간적 계층을 이용한 예측 방법을 제시하였다. 위 방법들은 단변량 시계열 자료가 가지고 있는 계층을 이용하여 정보량을 증가시키고 각 계층의 정보를 활용하여 예측 값을 조정하는 특징을 갖는다.

본 논문에서 소개할 다중 결합 예측 알고리즘(multiple aggregation prediction algorithm; MAPA)은 Kourentzes 등 (2014)이 제시한 방법으로 단변량 시계열 분석의 한계를 극복하는 분석 방법의 한 줄기인 시간적 결합을 이용한 방법이다. 다중 결합 예측 알고리즘은 시간적 결합의 방식과 앞서 제시한 예측값을 조합하는 방법을 결합한 예측 알고리즘으로 복수의 시간적 결합 자료를 생성한 후, 각각의 시간적 결합 자료에서 얻은 정보를 조합하여 최종 예측값을 계산하는 방식이다. 알고리즘의 자세한 내용은 2장에서 다루도록 한다. Spiliotis 등 (2018)은 다중 결합 예측 알고리즘을 활용하여 그리스 아테네의 은행 5개 지점의 전력 수요를 예측하였다.

본 논문은 다중 결합 예측 알고리즘을 이용하여 교통사고 발생 건수를 예측하고, 기존 시계열 분석에서 사용하는 일변량 예측 방식과 비교하였다. 국내 선행 연구에서는 Kim과 Lee (2014)는 ARIMA 모형을 사용하여 충청도 주요 도시의 교통사고 발생 건수를 예측하였고, Park 등 (2011)은 토지 이용 및 교통 특성을 반영하는 다중 회귀분석 모형을 사용하여 교통사고 예측 모형을 개발하였다. 이렇듯 대부분의 선행 연구는 일변량 시계열 자료만을 사용하는 예측과 일변량 시계열 자료와 외부인자의 회귀분석을 통한 예측에 국한되어 있다.

본 논문에서는 다중 결합 예측 알고리즘을 소개하고 실증 분석을 통하여 그 성능을 기존의 일변량 예측 방식과 비교한다. 논문은 총 4장으로 구성되어 있으며, 2장에서는 다중 결합 예측 알고리즘을 설명하고, 3장에서는 이를 이용하여 월별 국내 교통사고 발생 건수 자료를 예측한 후 그 성능을 비교하고, 4장에서는 결론을 제시한다.

2. 다중 결합 예측 알고리즘

본 장에서는 다중 결합 예측 알고리즘을 활용한 예측의 방법과 절차에 관해 설명하고자 한다. 다중 결합 예측 알고리즘은 일반적인 예측모형에서 중요하게 여겨지는 선택(selection)과 모수화(parameterisation)의 이슈의 중요성을 완화하는 장점이 있다. 알고리즘의 진행 절차는 다수의 시간적 결합 시계열 자료를 생성하고 각각의 결합 자료에서의 예측값을 조합하는 방법을 활용하여 예측 정확도를 향상시키는 알고리즘으로 총 3개의 과정으로 구분된다; (1) 시간적 결합 과정, (2) 예측 과정, 그리고 (3) 조합과정. 개괄적인 다중 결합 예측 알고리즘은 Figure 2.1에 그림으로 표현되어 있으며, 각 단계에 대한 설명은 2장에서 설명하도록 한다.

2.1. 시간적 결합 과정

시간적 결합이란 원 시계열 자료를 겹치지 않는 일정한 크기의 연속되는 집합으로 나누고, 각 집합의 평균을 사용하여 새로운 시계열 자료를 생성하는 방법이다. 집합의 크기를 변화시키면서 시간적 결합을

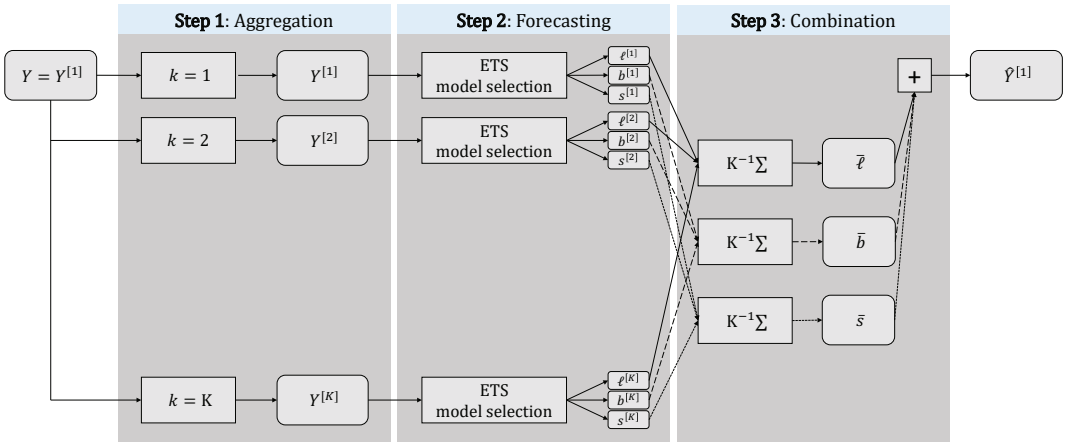

Figure 2.1. Overview of multi aggregation prediction algorithm. ETS = exponential smoothing method.

반복하며 하나의 시계열 자료로부터 여러 개의 시간적 결합 시계열 자료를 생성할 수 있다. 여기서 변화시키는 집합의 크기를 결합 수준(aggregation level)이라고 한다.

시간적 결합의 과정을 통해 원 시계열 자료가 가지고 있던 시계열 요소(time series component)들의 특성을 강화하거나 약화할 수 있다. 낮은 수준의 시간적 결합 데이터는 작은 결합 수준을 활용하여 만든 시간적 결합 자료로서 계절성이 뚜렷하게 나타나는 특징을 가진다. 또한, 가장 낮은 수준의 결합 데이터는 결합 수준으로 1을 사용하여 생성한 시간적 결합 자료로서 원 시계열 데이터와 동일한 데이터이다. 반면, 높은 수준의 시간적 결합 데이터는 큰 결합 수준을 활용하여 생성한 시간적 결합 자료로서, 대체로 수준(level)과 추세(trend) 시계열 요소가 강조되며 계절성은 약화되는 특징을 보인다. 시간적 결합 자료의 생성은 최대로 사용할 결합 수준을 K로 설정하고, 각각의 결합 수준에서 새로운 시계열 자료를 생성하는 방식으로 진행되며, 총 K개의 시계열 자료를 얻을 수 있다. 원 시계열 자료 Y가 $\{y_t : t = 1, 2, \ldots, n\}$라고 할 때, k 수준의 시간적 결합 자료 $Y^{[k]}$의 각 원소는 식 (2.1)과 같이 나타낼 수 있다.

$$y_i^{[k]} = k^{-1} \sum_{t=1+(i-1)k}^{ik} y_t, \quad i = 1, 2, \ldots, \lfloor n/k \rfloor. \tag{2.1}$$

식 (2.1)의 k는 시간적 결합의 수준(aggregation level), $\lfloor\ \rfloor$는 가우스 기호를 나타낸다. 시간적 결합 수준 k는 원 시계열 자료 개수인 n보다 작거나 같은 정수의 값을 가질 수 있다. 단, $Y^{[k]}$의 미래 예측값을 적절하게 추정하기 위한 샘플을 확보하기 위해서는 $k \ll n$의 k를 선호한다. k가 n의 약수가 아닌 경우 n/k의 값이 정수가 아니게 되며, 이러한 경우 n을 k로 나누었을 때의 나머지만큼의 자료를 원 시계열의 앞에서부터 제외한 후 시간적 결합 자료를 생성한다. Figure 2.2는 원 시계열 자료와 결합 수준에 따른 시간적 결합 자료의 그래프 예시이다. 원 시계열 자료는 총 36개의 값으로 구성되어 있다. 결합 수준을 4로 설정하는 경우, 원 시계열 자료의 개수 36이 결합 수준 4로 나누어떨어지므로 원 시계열에서 모든 자료를 사용하여 시간적 결합 자료를 생성할 수 있다. 반면 결합 수준을 5로 설정하는 경우, 원 시계열 자료의 개수 36이 결합 수준 5로 나누어지지 않으므로, 36을 5로 나눈 나머지인 1만큼의 시계열 자료를 앞에서 제외한 후 시간적 결합 자료를 생성한다. 이를 통해 $Y^{[k]}$의 미래 예측값의 첫 번째 값의 위치와 결합 수준이 1인 원 시계열 자료의 첫 번째 미래 예측값의 위치를 동일하게 설정할 수 있다.

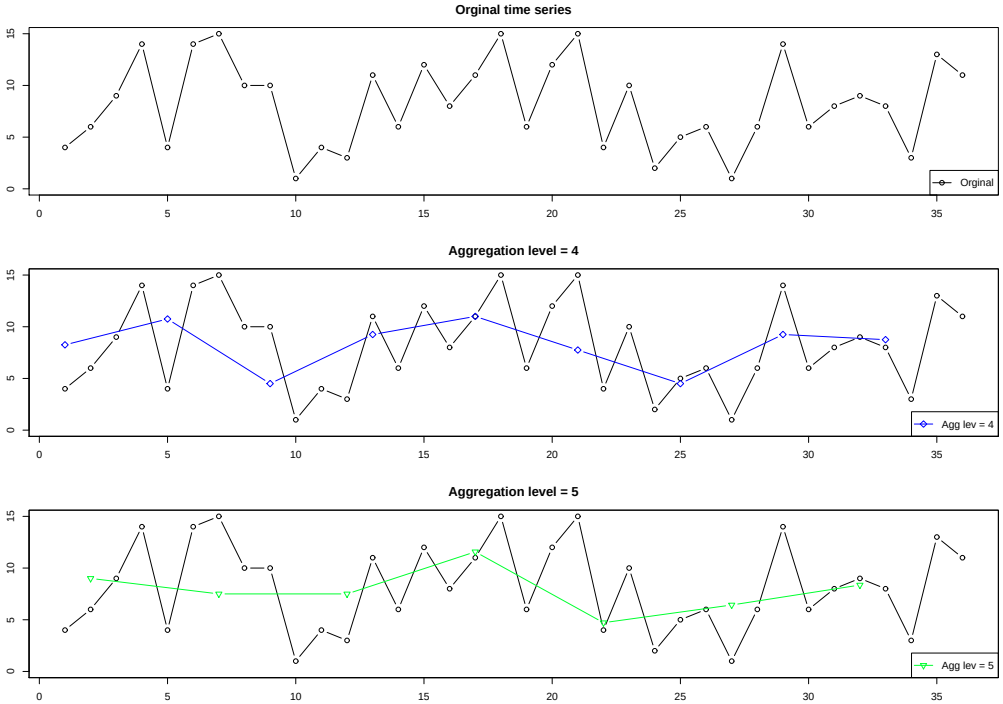

Figure 2.2. Temporal aggregation.

2.2. 예측 과정

시간적 결합 과정 단계를 통해 복수의 시간적 결합 자료를 얻었다면, 각각의 시간적 결합 자료에 알맞은 예측 모형을 적합해야 한다. 단, 적합된 예측 모형을 통해 예측값을 직접 계산하지 않고, 각각의 시간적 결합 자료에서 시계열 요소를 분해하는 과정을 추가한다. 이 과정을 통해 각각의 결합 수준에 대한 추세 요인과 계절 요인의 유무를 파악할 수 있다. 즉, 각 결합 수준이 가지고 있는 수준, 추세 또는 계절적 특성을 추출할 수 있다.

예측 모형을 적합하면서 시계열 요소를 분리해내는 과정에서 용이하게 사용할 수 있는 예측 방법은 지수평활법(exponential smoothing method; ETS)으로 Hyndman 등 (2008)이 정리한 예측 방법이다. 지수평활법은 수준 요인, 추세 요인, 계절 요인의 3가지 시계열 요소의 추출에 용이하므로 본 절에서 설명하는 예측 방식에 유용하게 사용할 수 있다. 그러므로 예측 모형의 적합 과정은 각 결합 수준의 시계열 데이터 $Y^{[k]}$에 각각 알맞은 ETS 모형을 적합하는 방식으로 진행된다. 예측값을 계산하는 과정에서 ETS 모형에서 얻을 수 있는 모든 정보를 사용하지는 않는다. 이후의 과정에서 필요한 요소는 최종 상태 벡터 $\mathbf{x}_i^{[k]} = (l_i, b_i, s_i, s_{i-1}, \ldots, s_{i-S^{[k]}+1})^T$이며 수준 요인은 l_i, 추세 요인은 b_i, 계절 요인은 s_i로 표현하였다. 그리고 $S^{[k]}$는 k 수준 데이터에서의 계절 주기를 의미한다.

추세 요인과 계절 요인이 결합 수준에 따라서 추정되거나 추정되지 않을 수 있다. 추세 요인의 경우, 특정한 시간적 결합 자료에 뚜렷한 추세가 보이지 않는 경우에는 추정되지 않을 수 있다. 또한, 계절 요인의 경우, 원 시계열 자료 $Y^{[1]}$가 m의 계절주기를 가진다면, 계절성은 m/k가 정수이고 $k < m$를 만족하는 k 수준 결합 시계열 자료에서만 나타날 수 있다. 이를 만족하는 경우, k 수준 시간적 결합 자료의 계절 주기는 m/k가 된다.

Table 2.1. Component prediction in the additve formulation

Trend	Seasonality		
	N	A	M
N	$l_{i+h} = l_i$	$l_{i+h} = l_i$ $s_{i-m+h} = s_{i-m+h}$	$l_{i+h} = l_i$ $s_{i-m+h} = (s_{i-m+h} - 1)l_{i+h}$
A	$l_{i+h} = l_i$ $b_{i+h} = hb_{i+h}$	$l_{i+h} = l_i$ $b_{i+h} = hb_{i+h}$ $s_{i-m+h} = s_{i-m+h}$	$l_{i+h} = l_i$ $b_{i+h} = hb_{i+h}$ $s_{i-m+h} = (s_{i-m+h} - 1)(l_{i+h} + b_{i+h})$
Ad	$l_{i+h} = l_i$ $b_{i+h} = \sum_{j=1}^{h} \phi^j b_i$	$l_{i+h} = l_i$ $b_{i+h} = \sum_{j=1}^{h} \phi^j b_i$ $s_{i-m+h} = s_{i-m+h}$	$l_{i+h} = l_i$ $b_{i+h} = \sum_{j=1}^{h} \phi^j b_i$ $s_{i-m+h} = (s_{i-m+h} - 1)(l_{i+h} + b_{i+h})$
M	$l_{i+h} = l_i$ $b_{i+h} = (b_i^h - 1)l_{i+h}$	$l_{i+h} = l_i$ $b_{i+h} = (b_i^h - 1)l_{i+h}$ $s_{i-m+h} = s_{i-m+h}$	$l_{i+h} = l_i$ $b_{i+h} = (b_i^h - 1)l_{i+h}$ $s_{i-m+h} = (s_{i-m+h} - 1)(l_{i+h} + b_{i+h})$
Md	$l_{i+h} = l_i$ $b_{i+h} = \left(b_i \sum_{j=1}^{h} \phi^j - 1\right) l_{i+h}$	$l_{i+h} = l_i$ $b_{i+h} = \left(b_i \sum_{j=1}^{h} \phi^j - 1\right) l_{i+h}$ $s_{i-m+h} = s_{i-m+h}$	$l_{i+h} = l_i$ $b_{i+h} = \left(b_i \sum_{j=1}^{h} \phi^j - 1\right) l_{i+h}$ $s_{i-m+h} = (s_{i-m+h} - 1)(l_{i+h} + b_{i+h})$

N = none; A = additive method; Ad = additive damped trend method; M = multiplicative method; Md = multiplicative damped trend method.

ETS 모형은 가법(additive)과 승법(multiplicative)의 두 가지 방법 중 하나를 사용하여 추세 요인과 계절 요인을 추정한다. 예를 들어, 결합 수준 $k = 2$의 데이터에서는 가법 방식으로 추세요인을 추정하고 결합 수준 $k = 6$의 자료에서는 승법 방식으로 추세요인을 추정할 수도 있다. 하지만 두 가지 방식은 수치상으로 직접적인 비교가 불가능하고 이후 조합 단계에서의 복잡성(complexity)을 증가시키기 때문에 승법 방식으로 추정된 요소를 가법 방식으로 변환하여 사용한다. 가법 형태로 변형된 요소들은 예측값을 만드는 과정에만 사용되며, 각각의 시계열 요소의 예측값은 Table 2.1의 식을 사용하여 계산한다. Table 2.1의 식은 기존 ETS 모형의 요소들의 예측식을 가법 형태로 변환한 표이며, 추세 요인이 M 혹은 Md이거나 계절 요인이 M인 경우가 가법 형태로 변환된 식으로 표현되어 있다. 여기서, M은 승법 모형을 나타내며 d는 약화 모형(damped model)을 나타낸다. 예를 들어, Md는 추세가 약화된 승법 모형을 의미한다.

각 결합 수준에서 가법 형태로 변환된 시계열 자료 $Y^{[k]}$의 예측값은 다음과 같이 계산된다.

$$\hat{y}_{i+h^{[k]}}^{[k]} = l_{i+h^{[k]}}^{[k]} + b_{i+h^{[k]}}^{[k]} + s_{i-m+h^{[k]}}^{[k]}. \tag{2.2}$$

식 (2.2)에서 $\hat{y}, l, b, s$의 $[k]$는 결합 수준을 의미하며 h의 $[k]$는 해당 위치의 결합 수준을 나타낸다. 이와 같은 표현은 같은 구간을 예측하더라도 결합 수준에 따라 예측값의 개수가 다르게 나타나는 특징을 반영하기 위해 사용한다. 예를 들어, 결합 수준 $k = 1$의 원 시계열 자료 $Y^{[1]}$에서 미래 24 시점을 예측하면 24개의 예측값이 생성되지만 같은 구간에서 $Y^{[12]}$의 예측값은 두 개밖에 생성되지 않는다. 이렇듯 개수의 차이가 있으므로 원 시계열의 시점에 모든 결합 수준에서의 예측값을 조합하기 위해서는 스케일을 원 시계열 자료의 스케일에 맞게 조정할 필요가 있다. 스케일 조정에는 Young (1967)이 제시하고 Kourentzes 등 (2014)이 다중 결합 예측 알고리즘에 적용한 piecewise constant interpolation 방법을 사용한다. 이 방법은 Figure 2.3에 그림을 통해 표현되어 있으며 3개의 결합 수준 $k = (1, 3, 4)$에서의 추세 요인의 예측값을 보여주고 있다. 여기서 결합수준 $k = (3, 4)$의 경우 $k = 1$의 스케일과 동일하게 조정할 필요가 있다. piecewise constant interpolation 방식은 스케일을 조정하기 위해 스케일 조정

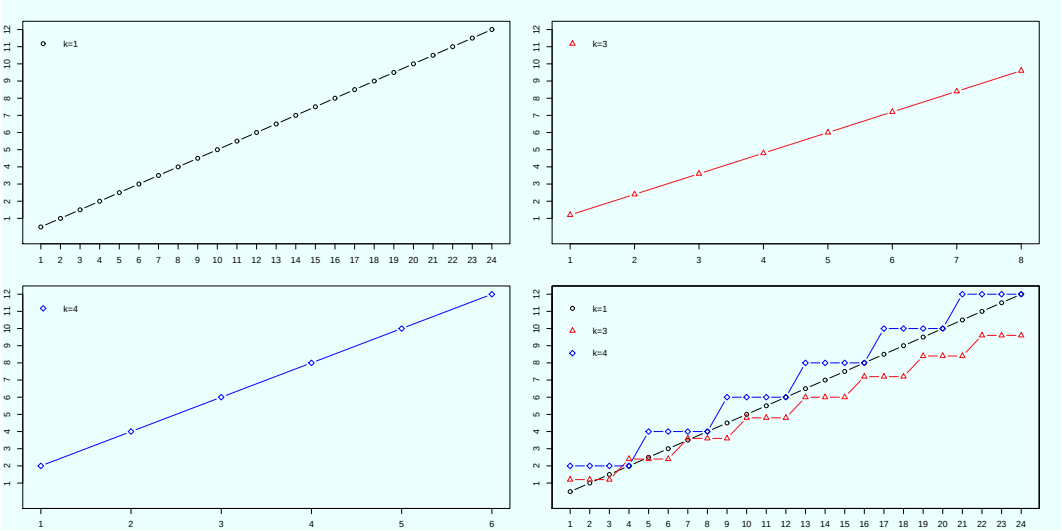

Figure 2.3. Estimated component for $k = (1, 3, 4)$ and in the original time scale.

이 필요한 결합 수준의 예측값을 결합 수준 만큼 반복하여 예측값이 비어있는 구간을 채우는 방식으로 진행한다. Figure 2.3의 마지막 그림에서 볼 수 있듯이, $k = 3$의 예측값은 3번씩 연속으로 반복되어 나타나고 $k = 4$의 예측값은 4번씩 연속으로 나타나면서 기존 데이터에 맞도록 스케일이 조정된 결과를 볼 수 있다.

단, ETS 모형이 항상 모든 시계열 요소를 포함하지는 않는다. 또한, 결합 수준에 따라 포함되는 시계열 요소가 다른 경우도 존재한다. 예를 들어, 계절 주기가 설정되지 않는 결합 수준의 데이터에서는 계절 요인이 추정되지 않으며, 추세 요인 또한 시계열 자료의 형태에 따라서 추정되지 않을 수 있다. 이러한 경우에는 각 요소의 예측값을 0으로 설정하고 이후 과정을 진행한다.

2.3. 조합 과정

다중 결합 예측 알고리즘의 마지막 단계는 3가지 시계열 요소의 예측값을 조합하는 과정으로, 시계열 요소의 예측값을 시계열 요소별로 구분하여 조합하는 단계이다. 각 시계열 요소 예측값의 조합은 산술 평균, 중위수, 가중 평균, 가중 중위수 등의 방식을 사용할 수 있다. Jose와 Winkler (2008)은 이와 같은 간단한 방식의 조합 과정을 통한 예측이 로버스트하고 좋은 예측력을 보여준다는 연구 결과를 제시하였다. 뿐만 아니라, 다중 결합 예측 알고리즘은 여러 결합 수준에서의 예측값을 조합하기 때문에 특히 장기 예측에서 좋은 결과를 보여준다고 알려져 있다.

이 절에서는 이후 실증 분석에서 사용할 방식인 산술 평균을 사용하는 조합 과정을 설명하려 한다. 산술 평균을 활용한 각 요소의 최종 예측값은 다음과 같이 정의된다.

$$\bar{l}_{t+h^{[1]}} = K^{-1} \sum_{k=1}^{K} l^{[k]}_{t+h^{[1]}}, \tag{2.3}$$

$$\bar{b}_{t+h^{[1]}} = K^{-1} \sum_{k=1}^{K} b^{[k]}_{t+h^{[1]}}, \tag{2.4}$$

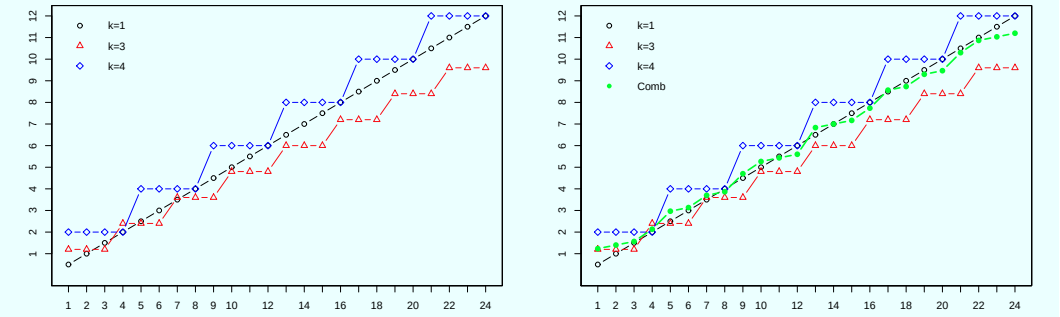

Figure 2.4. Combinated predicted value.

$$\bar{s}_{t+h^{[1]}} = K^{*-1} \sum_{k=1}^{K^*} s^{[k]}_{t+h^{[1]}}. \tag{2.5}$$

식 (2.5)의 K^*는 계절성이 발견되는 최대 결합 수준, m은 원 시계열 자료의 계절 주기를 나타내며 m/k가 정수이고 $k < m$를 만족하는 k의 경우에만 계산된다. Figure 2.4는 piecewise constant interpolation을 통해 얻은 추세 요인의 예측값과 산술평균을 통한 예측 조합 값의 예시 그래프이다. 왼쪽 그래프는 Figure 2.3의 마지막 그래프와 동일한 그림으로, piecewise constant interpolation이 완료된 $k = (1, 3, 4)$에서의 추세 요인의 예측값의 그래프이다. 오른쪽 그래프는 왼쪽 그래프에서 각 결합 수준의 예측값을 산술 평균으로 조합한 값이 추가된 그래프로 추세 요인이 평활된 형태로 나타나는 것을 확인할 수 있다. 최종적으로 원 시계열 자료의 스케일에 맞는 미래 시점의 예측값은 다음과 같이 표현할 수 있다.

$$\hat{y}^{[1]}_{t+h^{[1]}} = \bar{l}_{t+h^{[1]}} + \bar{b}_{t+h^{[1]}} + \bar{s}_{t-m+h^{[1]}}. \tag{2.6}$$

3. 실증분석

이 장에서는 실제 시계열 자료를 활용하여 시간적 결합을 이용한 예측을 시행한다. 다중 결합 예측 알고리즘을 활용하는 예측을 위해 R의 MAPA 패키지 (Kourentzes와 Petropoulos, 2018)의 함수인 mapa를 사용하였고, 각 결합 수준의 지수평활법 추정에는 R의 forecast 패키지 (Hyndman, 2019)의 함수인 ets 함수를 사용하였다. 각 결합 수준에서 최적의 지수평활법 모형을 추정하기 위한 기준으로 Burnham과 Anderson (2002)이 소개한 Akaike information criterion corrected (AICc) 정보량을 활용하였다.

3.1. 자료설명: 국내 교통사고 발생건수

실증 분석에 사용한 자료는 경찰 추산 국내 전국 교통사고 발생 건수이며, 2005년 1월부터 2018년 12월까지의 월별 자료이다. 총 시계열 자료의 길이는 168개이고 모형 적합을 위한 훈련 자료(training set)의 기간은 2005년 1월부터 2016년 12월까지 총 12년, 144 시점이다. 성능 평가를 위한 검증 자료(test set)의 기간은 2017년 1월부터 2018년 12월까지로 총 2년, 24 시점의 자료를 사용하였고 교통사고분석시스템(traffic accident analysis system, http://tass.koroad.or.kr)을 이용하여 얻을 수 있다.

Figure 3.1은 교통사고 발생 건수 자료의 월별 시계열 그래프와 연도별 월평균 시계열 그래프이다. 연

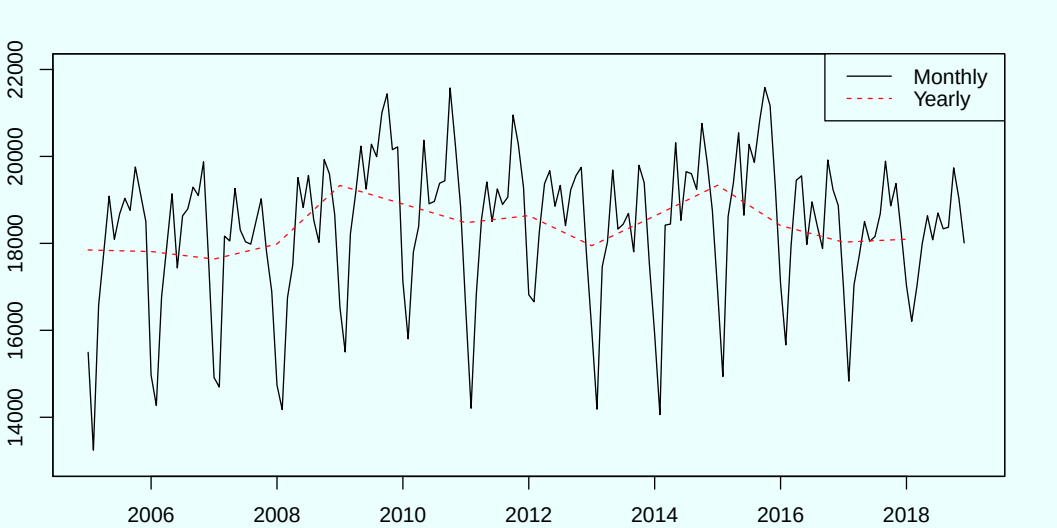

Figure 3.1. Time series plot of monthly and yearly data.

Table 3.1. Temporal aggregation of the traffic accident counts

Aggregation level	Length of time series
1	168
2	84
3	56
4	42
5	33
6	28
7	24
8	21
9	18
10	16
11	15
12	14

도별 월평균 그래프에서 볼 수 있듯이, 2009년까지는 증가하다가 이후로 감소하는 추세를 보인다. 하지만 2013년부터 2015년까지는 증가하고 다시 감소하는 추세를 보이는 등 총 14년의 구간 동안 명확하게 증가하거나 감소하는 추세가 보이지 않으며 전체적으로 일정한 수준에서 유지되는 것을 확인할 수 있다. 그리고 원 시계열 자료인 월별 시계열 그래프에서 알 수 있듯이 매년 2월에 교통사고 발생 건수가 최소이고, 대체로 10월에 교통사고 발생 건수가 최대인 12개월의 계절 주기를 가지는 것을 확인할 수 있다.

3.2. 시간적 결합

2장에서 설명한 다중 결합 예측 알고리즘을 활용하기 위하여 시간적 결합을 시행한다. 이 자료에서 활용한 시간적 결합 수준은 최대 12로, 각각의 시간적 결합 자료는 Table 3.1에 나타난 것과 같은 길이를 가지고 있다. 결합 수준 1은 원 시계열 자료와 동일하며 월별 전국 교통사고 발생 건수를 의미한다. 결합 수준 3은 3개월 단위로 나누어진 자료로 분기별 전국 교통사고 발생 건수의 월평균과 동일한 의미를 갖는다. 결합 수준 6의 경우 6개월 단위로 나누어진 자료로 반년별 교통사고 발생 건수의 월평균과 동

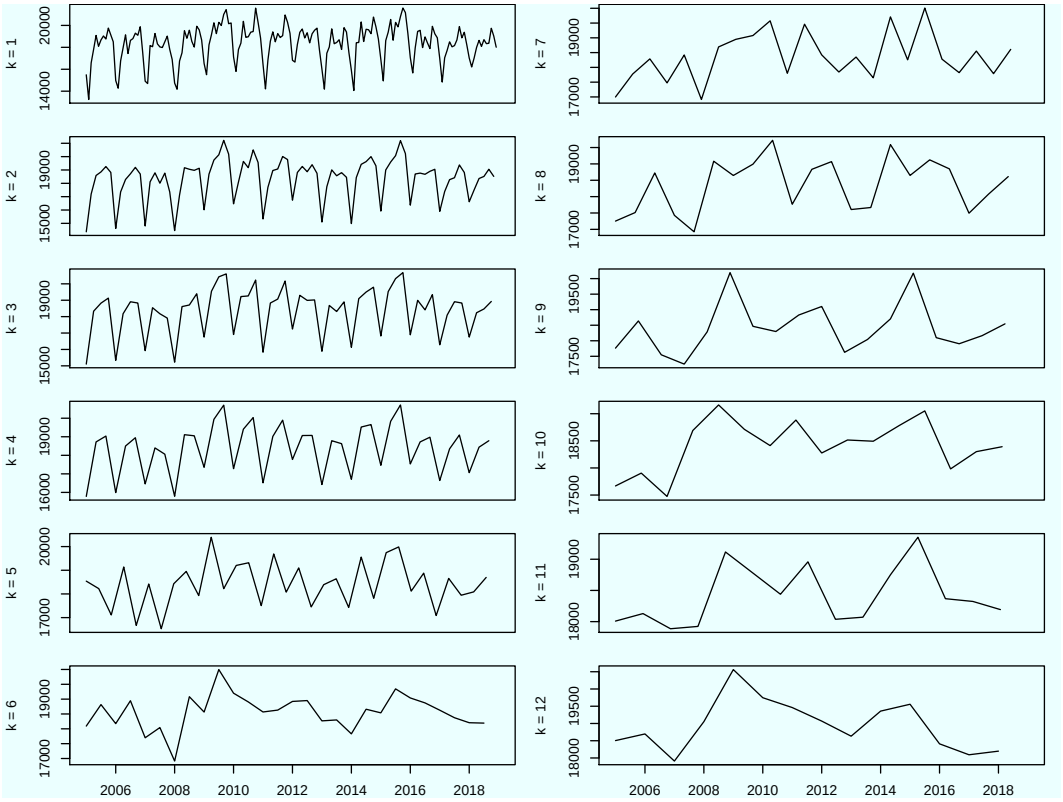

Figure 3.2. Time series plot of temporal aggregated series.

일하며, 결합 수준 12의 경우 연별 교통사고 발생 건수의 월평균과 동일하다.

기존 데이터의 계절 주기가 12이므로 2장 2절에서 제시한 계절 주기가 존재하는 조건인 $k < 12$이고 $12/k$가 정수라는 조건을 만족하는 결합 수준은 1, 2, 3, 4, 6이며 해당 결합 수준에서만 계절성이 존재할 수 있다. 이 결합 수준에 대응하는 계절주기는 12, 6, 4, 3, 2이다. Figure 3.2는 전체 데이터를 사용하여 만든 1부터 12까지의 결합 수준에서의 시계열 그림이다. 결합 수준 1의 원 시계열 자료에서는 계절성이 뚜렷하게 나타나며, 결합 수준 2와 3까지는 명확한 계절성이 보이는 것을 확인할 수 있다. 하지만 결합 수준 4와 6에서는 이전 결합수준에서 나타난 것처럼 계절성이 뚜렷하게 나타나지 않고 약해지는 모습을 확인할 수 있다.

3.3. 예측 비교

본 논문에서는 시계열 예측값에 대한 정확성 비교를 위하여 root mean squared error (RMSE), mean absolute percentage error (MAPE), 그리고 mean absolute scaled error (MASE)를 이용하였다. 원 시계열 자료 y_i와 예측값 $\hat{y}_i$에 대한 미래시점 h까지의 RMSE, MAPE, MASE는 아래처럼 정의된다.

$$\text{RMSE} = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(y_i - \hat{y}_i)^2}, \qquad \text{MAPE} = \frac{100}{h}\sum_{i=1}^{h}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \qquad \text{MASE} = \frac{1}{h}\sum_{i=1}^{h}\frac{|y_i - \hat{y}_i|}{Q}. \quad (3.1)$$

Table 3.2. Aggregation level selection

Maximum aggregation level	RMSE	MAPE	MASE
1	751.44	3.64	0.83
2	730.53	3.61	0.82
3	685.75	3.44	0.78
4	633.00	3.13	0.71
5	642.84	3.22	0.73
6	**589.98**	**2.97**	**0.67**
7	597.00	3.02	0.68
8	600.29	3.05	0.69
9	616.19	3.17	0.72
10	619.00	3.19	0.72
11	626.79	3.24	0.73
12	626.65	3.24	0.73

RMSE = root mean squared error; MAPE = mean absolute percentage error; MASE = mean absolute scaled error.

Table 3.3. Estimated ETS model according to aggregation level

Aggregation level	Model
1	ETS(A, N, A)
2	ETS(A, N, A)
3	ETS(A, N, A)
4	ETS(A, N, A)
5	ETS(A, N, N)
6	ETS(M, N, M)

The model column represents the taxonomy of exponential smoothing method as follow: ETS = method for estimating error, method for estimating trend, method for estimating seasonality. N = none; A = additive method; M = multiplicative method.

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

가장 적절한 최대 결합 수준을 선택하기 위해, 최대 결합 수준을 1에서 12까지로 변경하면서 산술 평균을 활용한 다중 결합 예측 알고리즘을 적용하였고, 검증 자료에서의 예측력이 가장 좋은 결합 수준을 선택하였다. Table 3.2에서 확인할 수 있듯이 결합 수준 $k = 6$일 때 검증 자료에서의 RMSE, MAPE, MASE가 가장 작게 나타나므로 해당 최대 결합 수준인 $k = 6$을 활용하는 예측값이 가장 정확한 것을 확인하였다. 이후 기존 모형과의 비교 단계에서 선택된 최대 결합 수준을 사용하였다.

최대 결합 수준을 6으로 설정했을 때, 각 결합 수준의 시계열 데이터에서 적합된 ETS 모형은 Table 3.3과 같이 나타난다. 각각의 ETS 모형은 R의 forecast 패키지의 ets 함수를 기반으로 자동으로 적절한 모형을 적합하였으며, 기준이 되는 정보량은 AICc이다. 계절 요인이 나타날 수 있는 결합 수준인 $k = (1, 2, 3, 4, 6)$에서 모두 계절 요인이 존재하는 것으로 추정되었다. 결합 수준 $k = 6$의 시간적 결합 데이터에서만 승법 방식으로 계절 요인이 추정되었으며 나머지 결합 수준에서는 가법 방식으로 추정되었다. 추세 요인의 경우 모든 결합 수준에서 추정되지 않았다. 이는 Figure 3.2에서 나타난 바와 같이 해당 결합 수준에서 뚜렷하게 증가하거나 감소하는 추세가 보이지 않기 때문에 추세 요인이 존재하지 않는다고 추정된 것을 알 수 있다.

알고리즘을 활용한 예측의 성능을 비교하기 위해, 대조 모형으로 원 시계열 데이터에 대한 ETS 모형과 ARIMA 모형을 사용하였다. ETS 모형과 ARIMA 모형은 R의 forecast 패키지의 ets 함수

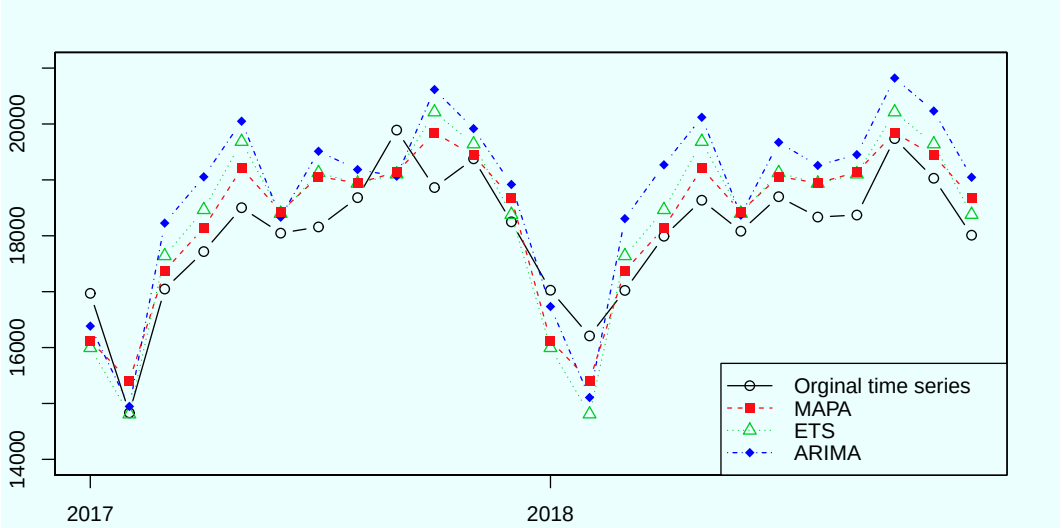

Figure 3.3. Comparison. MAPA = multiple aggregation prediction algorithm; ETS = exponential smoothing method; ARIMA = autoregressive integrated moving averag.

Table 3.4. RMSE, MAPE, and MASE for test set

	RMSE	MAPE	MASE
ETS(A, N, A)	1043.12	5.21	1.20
ARIMA$(2,0,0)(2,1,0)[12]$	751.44	3.64	0.83
MAPA(agg.lev = 6)	**589.98**	**2.97**	**0.43**

RMSE = root mean squared error; MAPE = mean absolute percentage error; MASE = mean absolute scaled error; ETS = exponential smoothing method; ARIMA = autoregressive integrated moving averag; N = none; A = additive method.

와 auto.arima 함수를 활용하여 추정하였다. AICc 정보량을 기준으로 선택된 ETS 모형은 ETS(A, N, A)이며, 이는 결합 수준이 1인 시간적 결합 자료에서 적합된 모형과 동일하다. ARIMA 모형은 ARIMA$(2,0,0)(2,1,0)[12]$이 선택되었다.

검증 자료와 다중 결합 예측 알고리즘, ETS(A, N, A), 그리고 ARIMA$(2,0,0)(2,1,0)[12]$의 예측값에 대한 그래프는 Figure 3.3에서 볼 수 있으며 RMSE, MAPE, 그리고 MASE를 활용한 정확성 비교는 Table 3.4에서 확인할 수 있다. Figure 3.3에서 확인할 수 있듯이, 2년의 구간 동안 다중 결합 예측 알고리즘의 예측값이 실제 검증 자료의 값이 대부분 시점에서 가깝게 나타나는 것을 확인할 수 있다. 다만, 세 가지 방법 모두 2년의 예측 구간에서 실제보다 큰 값으로 추정된 것을 확인할 수 있다. 이는 과거의 시계열 수준이 예측 구간인 2017년과 2018년에 비해 상대적으로 큰 값을 가지고 이것이 예측에 반영된 결과이다. 또한, Table 3.4에서도 확인할 수 있듯이 RMSE, MAPE, MASE의 모든 예측 정확성 측도에서, 다중 결합 예측 방법은 다른 모든 모형과 비교하여 가장 낮은 값을 보여주며 좋은 예측 성능을 가지고 있는 것으로 확인된다.

본 논문에서는 시간적 결합 과정을 집합이 서로 겹치지 않는(non-overlapping) 방식으로 진행하였다. 본 논문에서 자세히 다루지는 않았지만, 시간적 결합 과정에서 집합이 서로 겹치는(overlapping) 방식의 실험 또한 진행하였다. 시간적 결합 과정만 변형한 다중 결합 예측 알고리즘 실증 분석 결과는 기존의 다중 결합 예측 알고리즘과 같이 향상된 결과를 보이지 않고, 도리어 예측력이 알고리즘의 기초가 되는 단변량 ETS 모형보다도 감소하는 결과를 보이기도 하였다. 이는 계절성이 매우 뚜렷하게 나타나는

Table 3.5. Time series cross validation with non-fixed aggregation level and models

Forecast horizon	ETS	ARIMA	MAPA	Best
1	756.18	755.21	**743.46**	**MAPA**
2	824.68	781.42	**629.47**	**MAPA**
3	875.05	789.96	**618.28**	**MAPA**
4	873.59	797.52	**622.87**	**MAPA**
5	795.21	658.78	**630.37**	**MAPA**
6	896.95	641.74	**610.56**	**MAPA**
7	924.06	**622.19**	690.60	ARIMA
8	944.35	644.87	**580.85**	**MAPA**
9	971.98	726.02	**591.31**	**MAPA**
10	822.00	730.16	**480.72**	**MAPA**
11	606.80	591.85	**362.09**	**MAPA**
12	466.49	606.24	**326.77**	**MAPA**

ETS = exponential smoothing method; ARIMA = autoregressive integrated moving averag; MAPA = multiple aggregation prediction algorithm.

시계열 자료에서 서로 겹치는 방식의 시간적 결합을 사용하는 경우 나타나는 결과이다. 서로 겹치는 방식의 시간적 결합은 이동 평균(moving average) 방식의 분해법과 비슷한 방법으로, 모든 결합 수준에서 하나의 단위 구간 내에 포함되는 시점의 수가 동일하게 되며 결합 수준이 증가할수록 집합이 겹치지 않는 시간적 결합 자료에 비해 자료가 빠르게 평활되어 계절 요인이 추정되지 않는다. 이는 계절 요인의 예측값을 조합하는 과정에서 평균에 포함되는 0의 개수를 증가시키게 되어 원 시계열 자료의 계절성을 제대로 반영하지 못하는 결과를 낳는다. 다만, Boylan과 Babai (2016)은 집합이 서로 겹치는 시간적 결합 방식이 집합이 서로 겹치지 않는 시간적 결합 방식의 성능을 비교하였고 간헐적(intermittent) 자료의 경우 집합이 겹치는 시간적 결합 방식이 향상된 결과를 보인다는 연구 결과가 있기 때문에 추가적인 연구가 필요하다고 하겠다.

3.4. 시계열 교차검증

추가로 시계열 교차검증의 방식을 활용하여 다중 결합 예측 알고리즘의 정확성을 평가하였다. 기본적으로 훈련세트로 2016년 12월까지의 자료를 활용하였으며, 2017년 이후의 시점을 하나씩 추가하며 예측 정확성을 평가하였고, 최대 예측 시점은 12 시점으로 설정하였다. 비교 대상으로는 동일하게 ETS 모형과 ARIMA 모형을 사용하였다. 예측 정확성을 평가하는 기준은 RMSE를 사용하여 비교하였다.

먼저, 단계별로 결합 수준 혹은 모형을 각 단계의 훈련 세트에서 가장 좋은 정확성 및 정보량을 갖는 결합 수준 및 모형으로 설정하는 방식을 활용하여 비교하였다. 자세히 설명하면, 다중 결합 예측 알고리즘은 12 시점 이후까지의 예측값을 활용하여 MASE를 계산하고 MASE가 가장 작게 나타나는 결합 수준을 활용하여 예측값을 계산하였다. ETS 모형과 ARIMA 모형은 훈련 세트에서 AICc시 최소로 나타나는 모형을 활용하여 예측값을 계산하여 비교하였다. 즉, 한 시점이 추가될 때마다 가장 적합한 결합 수준 및 모형을 변경하였다. 위 과정을 통하여 얻은 교차검증 RMSE는 Table 3.5와 같이 계산되었다.

고정되지 않는 결합 시점 및 모형을 활용한 비교에서는 7 시점 미래를 예측하는 경우에만 ARIMA 모형이 가장 정확한 결과를 보여주었고, 이를 제외한 모든 경우에서 다중 결합 예측 알고리즘의 RMSE가 가장 작게 나타나며 예측 정확도가 높은 것을 보여준다. 그리고 10, 11, 12 시점 미래 예측과 같이 먼 미래의 값을 예측하는 경우의 정확도가 비교 대상에 비해 상대적으로 더 높은 정확성을 보이는 것을 확인할

Table 3.6. Time series cross validation with fixed aggregation level and models

Forecast horizon	ETS	ARIMA	MAPA	Best
1	756.18	839.71	**748.63**	**MAPA**
2	824.68	880.48	**636.27**	**MAPA**
3	875.05	896.74	**614.65**	**MAPA**
4	873.59	883.15	**621.58**	**MAPA**
5	795.21	793.76	**643.36**	**MAPA**
6	896.95	816.47	**618.35**	**MAPA**
7	924.06	829.21	**699.43**	**MAPA**
8	944.35	802.63	**590.33**	**MAPA**
9	971.98	854.38	**595.88**	**MAPA**
10	822.00	837.50	**499.72**	**MAPA**
11	606.80	709.06	**374.14**	**MAPA**
12	466.49	709.79	**354.56**	**MAPA**

ETS = exponential smoothing method; ARIMA = autoregressive integrated moving averag; MAPA = multiple aggregation prediction algorithm.

수 있다.

다른 시계열 교차검증 방법으로는 3장 3절에서 선택된 최적 결합 수준인 $k = 6$과 최적 모형으로 선택된 ETS(A, N, A)와 ARIMA$(2, 0, 0)(2, 1, 1)$[12] 모형을 변경하지 않고 활용하여 동일한 방법으로 시계열 교차검증을 시행하였다. 시계열 교차검증을 통해 계산된 RMSE는 Table 3.6과 같이 나타난다. 모든 예측 시점에서 다중 결합 예측 알고리즘이 가장 낮은 RMSE를 보이기 때문에, 해당 방식이 가장 정확한 방법인 것으로 나타났다. 고정된 모형을 활용한 교차검증에서도 먼 미래의 값을 예측하는 경우가 비교 대상에 비해 상대적으로 높은 정확성을 보인다.

4. 결론

본 논문에서는 일변량 시계열 자료를 이용하여 복수의 시간적 결합 자료를 만들고, 각각의 시간적 결합 자료를 ETS 모형을 통해 시계열 요소를 분리한 후, 각 요소의 예측값을 조합하여 예측 성능을 높이는 방법인 다중 결합 예측 알고리즘을 소개하고 있다. 실증 분석으로 국내 월별 교통사고 발생 건수 자료를 이용하여 여러 개의 시간적 결합 자료를 만든 후, 지수평활법을 사용하여 각 결합 수준의 시계열 요소를 분리하고, 분리한 요소들을 다시 조합하여 예측하는 방법의 장점을 살펴보았다. 기존 일변량 시계열 분석에서 사용하는 ETS 모형과 ARIMA 모형과 비교하여 다중 결합 예측 알고리즘을 우수한 성능을 보이는 것을 확인할 수 있었다. 시계열 교차검증 결과 장기 예측 부분에서 월등한 성능을 보이는 것을 확인할 수 있었다. 다만, 실증 분석에 사용하지 않은 자료에서는 실증 분석 결과와는 다르게 다중 결합 예측 알고리즘이 알고리즘의 기초가 되는 ETS 모형보다는 좋은 성능을 보이지만 ARIMA 모형보다는 우수하지 못한 결과를 보이는 경우가 있으므로 이러한 문제를 개선해 나갈 필요가 있다. 또한, 다중 결합 예측 알고리즘은 시계열 요소의 분리를 위해 지수평활법만을 활용하였다. 시계열 요소를 분리하는 방법에는 Cleveland 등 (1990)이 소개한 seasonal and trend decomposition using Loess (STL) 및 Dagum과 Bianconcini (2016)이 제시한 seasonal extraction in ARIMA time series (SEATS) 등과 같은 방법도 있으므로 이 방식을 사용하여 예측 정확성을 높일 수 있는 연구가 필요하다. 본 논문에서는 하나의 계절 주기를 가지는 자료를 활용하였지만, 자료 수집의 단위가 작아지고 여러 개의 계절 주기를 포함하는 자료가 많아지고 있기 때문에 다중 계절성을 가지는 자료에서도 이와 같은 알고리즘을 활용할

수 있도록 하는 연구가 필요하다. 그뿐만 아니라, 최적의 최대 결합 수준을 설정하는 방법에 관한 연구 또한 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

Boylan, J. E. and Babai, M. Z. (2016). On the performance of overlapping and non-overlapping temporal demand aggregation approaches, *International Journal of Production Economics*, **181(A)**, 136–144.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed), Springer-Verlag.

Casals, J., Jerez, M., and Sotoca, S. (2009). Modelling and forecasting time series sampled at different frequencies, *Journal of Forecasting*, **28**, 316–342.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess, *Journal of Official Statistics*, **6**, 3–33.

Dagum, E. B. and Bianconcini, S. (2016). *Seasonal Adjustment Methods and Real Time Trend-Cycle Estimation*, Springer.

Hyndman, R. J. (2019). Forecast: forecasting functions for time series and linear models. R package version 8.7. http://pkg.robjhyndman.com/forecast/

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., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*, Springer-Verlag, Berlin.

Jose, V. R. R. and Winkler, R. L. (2008). Simple robust averages of forecasts: some empirical results, *International Journal of Forecasting*, **24**, 163–169.

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

Kourentzes, N. and Petropoulos, F. (2018). MAPA: Multiple Aggregation Prediction Algorithm. R package version 2.0.4. https://github.com/trnnick/mapa/

Kourentzes, N., Petropoulos, F., and Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies, *International Journal of Forecasting*, **30**, 291–302.

Park, J., Jang, I., Son, E., and Lee, S. (2011). Development of traffic accident forecasting models considering urban-transportation system characteristics, *Journal of Korean Society of Transportation*, **29**, 39–56.

Rossana, R. J. and Seater, J. J. (1995). Temporal aggregation and economic time series, *Journal of Business & Economic Statistics*, **13**, 441–451.

Rostami-Tabar, B., Babai, M. Z., Syntetos, A., and Ducq, Y. (2013). Demand forecasting by temporal aggregation, *Naval Research Logistics (NRL)*, **61**, 489–500.

Spiliotis, E., Petropoulos, F., Kourentzes, N., and Assimakopoulos, V. (2018). Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption, Technical Report, National Technical University of Athens, Athens.

Trabelsi, A. and Hillmer, S. (1989). A benchmarking approach to forecast combination, *Journal of Business and Economic Statistics*, **7**, 353–362.

Young, S, W. (1967). Piecewise monotone polynomial interpolation, *Bulletin of the American Mathmatical Society*, **73**, 642–643.

다중 결합 예측 알고리즘을 이용한 교통사고 발생건수 예측

배두람[a] · 성병찬[a,1]

[a]중앙대학교 응용통계학과

(2019년 8월 6일 접수, 2019년 9월 20일 수정, 2019년 10월 4일 채택)

요 약

하나의 시계열 자료에서 다양한 특징을 발견하는 일은 간단한 문제가 아니다. 본 논문에서는 하나의 시계열 자료에서 복수의 패턴을 찾아내어 예측 정확도를 높이는 방식인 다중 결합 예측 알고리즘을 소개한다. 이 알고리즘은 시간적 결합과 예측값 조합의 개념을 사용한다. 시간적 결합 방식을 통해, 하나의 시계열 자료에서 여러 개의 시계열 자료를 생성할 수 있으며, 각각의 자료는 별도의 특성을 가지게 된다. 여러 개의 시계열 자료에서 다양한 특성을 추출하기 위하여 지수평활법을 사용하고 시계열 요소들 및 이들의 예측값을 계산한다. 마지막 단계에서 시계열 요소 별로 예측값을 혼합 한 후, 각 시계열 요소들의 조합값을 더하여 최종 예측값을 만든다. 실증 분석으로 국내 교통사고 발생 건수를 예측한다. 분석 결과, 기존의 다른 예측 방식보다 예측 성능이 우수함을 확인할 수 있다.

주요용어: 시간적 결합, 예측 조합, 다중 결합 예측 알고리즘, 시계열 요소, 지수 평활법

이 논문은 2018년도 중앙대학교 CAU GRS 지원에 의하여 작성되었음.

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