The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.1.181
30(1), 181–193

Hierarchical time series forecasting with an application to traffic accident counts

Jooeun Lee[a] · Byeongchan Seong[a,1]

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

(Received January 5, 2017; Revised January 16, 2017; Accepted January 23, 2017)

Abstract

The paper introduces bottom-up and optimal combination methods that can analyze and forecast hierarchical time series. These methods allow forecasts at lower levels to be summed consistently to upper levels without any ad-hoc adjustment. They can also potentially improve forecast performance in comparison to independent forecasts. We forecast regional traffic accident counts as time series data in order to identify efficiency gains from hierarchical forecasting. We observe that bottom-up or optimal combination methods are superior to independent methods in terms of forecast accuracy.

Keywords: grouped time series, revised forecasts, bottom-up forecasts, optimal combination forecasts, ARIMA model, exponential smoothing method

1. 서론

최근 데이터의 수집과 저장 기술의 발전으로 구조적으로 계층화 또는 집단화되어 있는 다중 시계열(multiple time series) 분석에 대한 관심이 증가되고 있다. 이런 다중 시계열이 성별, 지리, 또는 제품 종류와 같은 범위에 따라 계층적 또는 그룹적 구조를 가지고 있다는 점은, 계층적 시계열을 어떻게 모형화하고 예측할 것인가에 대한 문제로 이어진다.

예측 분야에서 계층적 또는 그룹화된 시계열(hierarchical or grouped time series)의 분석은 비교적 최근에 관심을 받게 된 주제로서 다양한 분야에서 응용되고 있다 (Athanasopoulos 등, 2009). 거시 경제학의 예측에서는 Weale (1988)이 국가경제계정(national economic account)을 생산, 수입과 지출, 자본 거래로 분해하였다. 생산은 각 국가에 대한 생산으로 분류되며, 수입과 지출 및 자본 거래는 개인, 회사, 정부 기관과 기타로 분류된다. 이 예는 분해되는 순서가 위에서 아래로 유일하다는 점에서 계층적 시계열로 볼 수 있다. 한편, 인구통계학의 예측(demographic forecasting)을 예로 들 때, 사망자 수는 성별, 그리고 지역별 사망자 수로 분류할 수 있다. 이는 유일한 계층 구조를 가지고 있지 않은 계층적 시계열 모형으로서 그룹화된 시계열의 예로 볼 수 있다. 즉, 특정 국가의 사망자 수는 지역별 또는 성별 분류의 상하 분해 순서가 정해져 있지 않다.

계층적 시계열 예측에 대한 논의는 주로 상향식(bottom-up) 방법과 최적조합(optimal combination)

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2015.
[1]Corresponding author: Department of Applied Statistics, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul 06974, Korea. E-mail: bcseong@cau.ac.kr

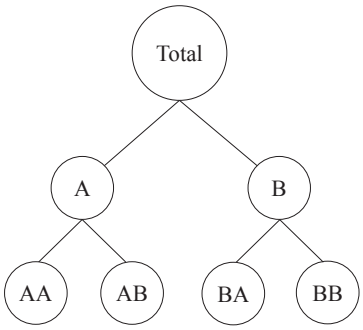

Figure 2.1. Two-level hierarchical structure.

방법으로 진행된다 (Shang과 Smith, 2013). 상향식 방법은 최하위 계층의 시계열을 예측한 후 이를 위로 합하면서 높은 단계의 예측을 얻어내는 방법이다. Hyndman 등 (2011)이 제안한 최적조합 방법은 예측을 정해진 방향 없이 선형회귀모형을 기반으로 최소제곱추정량(ordinary least square estimator)을 사용하여 계층화된 예측값을 계산한다. 이러한 계층적 시계열 예측 방법을 사용하면 임의의 조정 과정을 거치지 않고도 하위 계층의 예측값의 합은 항상 상위 계층의 예측값과 일치하게 된다. 따라서 임의의 조정 과정이 필요한 개별적 또는 독립적 예측과 비교할 때 예측 정확도를 높일 수 있다.

본 논문에서는 계층적 시계열 분석법으로 교통사고 발생건수를 예측하고 단변량 시계열 예측법과 비교한다. 국내의 사전 연구들은 주로 단변량 모형에 의존하여 교통사고 관련 시계열 자료를 분석 및 예측하였다. Han과 Kim (2007)은 ARIMA 모형을 사용하여 도로 종류별 사망자 수를 추세 모형화하고 향후 5년의 교통사고 사망자 수를 예측하였으며, Kim과 Lee (2014)는 ARIMA 모형을 이용한 교통사고 건수 예측치의 적정성을 검토하였다. Han (2007)은 회귀분석 모형과 시계열분석 모형으로 특정 시도의 교통사고 발생 추이를 파악하여 현재의 교통사고 추세가 중장기적으로 어떻게 변화할 것인지 추정하였다. 그러나, 이러한 모든 국내의 연구들은 단변량 또는 외생적 시계열이 특정 단변량 시계열에 미치는 영향을 연구하는 것에 제한되어 있다.

본 논문에서는 계층적 시계열 자료 분석을 위한 대표적인 두 가지 방법을 소개하고 실증 분석을 통하여 그 효율성을 독립적 예측과 비교한다. 본 논문은 총 4장으로 구성되어 있으며, 2장에서 상향식 및 최적조합 방법에 의한 계층적 예측을 설명하고 3장에서는 이를 이용하여 국내 교통사고 발생건수 시계열 자료를 예측하였다. 마지막 4장에서는 결론을 맺는다.

2. 계층적 시계열 예측 방법

2.1. 계층적 시계열

계층적 시계열은 다단계의 계층을 가지고 있으며, 최상위 단계(top-level or level 0)는 완전히 전체가 통합된 계열(completely aggregated series, 완전통합계열)을 나타내며, 1-단계(level 1)는 첫 단계의 분해(disaggregation)를 나타낸다. 예를 들어, Figure 2.1은 두 단계의 계층적 시계열 구조를 도식화하여 표현한 것으로, A는 1-단계에서의 A계열을 나타내고 AB는 1-단계의 A계열에 속한 2-단계에서의 B계열을 나타낸다.

Figure 2.1을 구성하고 있는 계층적 시계열의 구조는 다음과 같이 나타낼 수 있다.

$$Y_{\text{Total},t} = Y_{A,t} + Y_{B,t}, \quad Y_{A,t} = Y_{AA,t} + Y_{AB,t}, \quad Y_{B,t} = Y_{BA,t} + Y_{BB,t}. \tag{2.1}$$

즉, 상위 단계의 관측값은 하위 단계 계열의 합으로 얻어진다. 일반적으로, 계층화 시계열 벡터 $\boldsymbol{Y}_t$는 다음과 같이 나타낸다.

$$\boldsymbol{Y}_t = \left[Y_t, \boldsymbol{Y}'_{1,t}, \ldots, \boldsymbol{Y}'_{K,t}\right]', \tag{2.2}$$

여기서, Y_t $(t = 1, \ldots, T)$는 Figure 2.1의 $Y_{\text{Total},t}$처럼 완전통합계열을 나타내는 단변량 시계열이며, $\boldsymbol{Y}_{k,t}$ $(k = 1, \ldots, K)$는 시간 t에서 k-단계에 해당하는 모든 관측값들로 이루어진 벡터이며 K는 단계(level)의 총 개수 또는 최하위 단계(bottom-level)를 나타낸다. 예를 들어, Figure 2.1에서 $K = 2$이다. $\boldsymbol{Y}_t$는 모든 계층의 계열을 포함하고 있다. 계층적 시계열의 구조는 다음의 식으로 표현할 수 있다.

$$\boldsymbol{Y}_t = \boldsymbol{S}\boldsymbol{Y}_{K,t}. \tag{2.3}$$

식 (2.3)은 최하위 단계의 계열인 $\boldsymbol{Y}_{K,t}$와 모든 계층의 계열 $\boldsymbol{Y}_t$를 행렬 $\boldsymbol{S}$로 연결하고 있다. 행렬 $\boldsymbol{S}$는 $(m \times m_K)$차원의 행렬로서 합계 행렬(summing matrix)이라고 부르며, $m = 1 + m_1 + \cdots + m_K$은 전체 계열의 수이고, m_k는 k-단계에 해당하는 총 계열의 수를 나타낸다. 예를 들면, Figure 2.1에서 $m = 1 + 2^1 + 2^2 = 7$이다. 또한, 합계 행렬 $\boldsymbol{S}$를 이용하여 Figure 2.1의 계층을 다음과 같이 표현할 수 있다.

$$\boldsymbol{Y}_t = \begin{bmatrix} Y_t \\ Y_{A,t} \\ Y_{B,t} \\ Y_{AA,t} \\ Y_{AB,t} \\ Y_{BA,t} \\ Y_{BB,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Y_{AA,t} \\ Y_{AB,t} \\ Y_{BA,t} \\ Y_{BB,t} \end{bmatrix} (= \boldsymbol{S}\boldsymbol{Y}_{K,t}). \tag{2.4}$$

계층적 시계열의 예측은, 시간 T까지 주어진 정보를 바탕으로 미래 시점에 대한 계층적 시계열을 구성하는 모든 m개의 계열에 대한 개별적인 예측을 수행함을 의미한다. 예측 표기법으로 $\widehat{Y}_{0,h}$는 완전통합계열에 대한 미래 h-시점(h-step-ahead) 예측을, $\widehat{Y}_{A,h}$는 A계열에 대한 미래 h-시점 예측을, $\widehat{Y}_{AA,h}$는 AA계열에 대한 미래 h-시점 예측을 나타낸다. 이러한 계층적 계열들에 대한 예측의 초기값은 일반적인 단변량 시계열을 위한 ARIMA 모형 또는 지수평활법(exponential smoothing method)과 같은 방법을 이용하여 생성할 수 있다. 이와 같이 계층적 시계열의 초기 예측을 기저 예측(base forecast)이라고 부른다 (Hyndman과 Athanasopoulos, 2014). 기저 예측은 계층 구조가 전혀 반영되지 않은 것이므로, 여기에 계층 구조를 반영하면 최종 예측값을 만들 수 있다. 이것을 수정 예측(revised forecast)라고 하고, $\tilde{Y}_{0,h}$ 및 단계 $k = 1, \ldots, K$에 대하여 $\tilde{\boldsymbol{Y}}_{k,h}$와 같이 표현한다.

수정 예측을 얻기 위한 계층적 시계열 예측은 하향식(top-down) 방법, 상향식 방법, 그리고 이들을 조합하는 방법(예를 들면, middle-out 방법) 등 여러 통합 방법이 존재한다. 최근, 계층 간에 존재하는 임의의 상관관계를 고려하기 위하여 Hyndman 등 (2011)은 최적조합 예측을 제안하였다. 본 논문에서는 상향식 방법과 최적조합 방법의 두 가지 접근법을 다룬다.

2.2. 상향식 방법

가장 자주 사용하는 계층적 예측 방법은 상향식 방법이다 (Zellner와 Tobias, 2000). 이 방법은 먼저 독립적으로 최하위 단계의 각 계열에 대해 예측을 수행한 후, 전체 계층에 대한 수정 예측을 얻기 위해 위

로 합쳐 나간다. Figure 2.1과 같은 계층을 예로 들면, 최하위 단계 계열에 대해 독립적인 미래 h-시점 예측을 생성한 후, 이를 각각 $\widehat{Y}_{AA,h}$, $\widehat{Y}_{AB,h}$, $\widehat{Y}_{BA,h}$, $\widehat{Y}_{BB,h}$라고 하자. 이를 상위 단계의 계층으로 통합함으로써 나머지 계열에 대한 미래 h-시점 예측을 얻을 수 있다.

$$\bar{Y}_{A,h} = \bar{Y}_{AA,h} + \bar{Y}_{AB,h}, \quad \bar{Y}_{B,h} = \bar{Y}_{BA,h} + \bar{Y}_{BB,h}, \quad \bar{Y}_{h} = \bar{Y}_{A,h} + \bar{Y}_{B,h}.$$

상향식 방법에서는 최하위 단계 계열에 대한 수정 예측은 기저 예측과 동일하다. 이것은 상향식 방법의 고유 성질로서 ARIMA 모형 또는 지수평활법 사용과 관계가 없다(예를 들면, $\bar{Y}_{AA,h} = \widehat{Y}_{AA,h}$). 따라서, 상향식 방법을 합계 행렬을 이용하여 다음과 같이 나타낼 수 있다.

$$\bar{\boldsymbol{Y}}_h = \boldsymbol{S}\widehat{\boldsymbol{Y}}_{K,h}, \tag{2.5}$$

여기서, $\bar{\boldsymbol{Y}}_h = [\bar{Y}_{0,h}, \bar{\boldsymbol{Y}}'_{1,h}, \ldots, \bar{\boldsymbol{Y}}'_{K,h}]'$는 모든 계층에 대한 수정 예측이며, $\widehat{\boldsymbol{Y}}_{K,h}$는 최하위 단계($K$-단계)의 예측을 의미한다. 상향식 방법의 장점은 통합 과정에서 어떤 정보의 손실도 발생하지 않는다는 점이다. 하지만, 최하위 단계에서 이상점 또는 편차가 존재하는 경우, 이러한 효과가 최상위 계열의 부정확한 예측값과 연결될 수 있다.

2.3. 최적조합 방법

최적조합 방법은 상향식 방법과 달리 모든 계층적 시계열에 대해 기저 예측을 생성한다. 그러나, 이러한 기저 예측은 독립적으로 생성되기 때문에 계층 구조가 전혀 반영되어 있지 않다. 최적조합 방법은 이러한 기저 예측을 계층 구조에 따라 선형회귀모형으로 통합하고 단변량 예측과 최대한 가깝도록 조정하는 역할을 한다. 다른 방법과 달리 계층 내 모든 가능한 정보를 사용하며, 각 단계의 계열 간 상관관계와 상호작용을 반영한다고 알려져 있다. 또한 기저 예측이 불편성(unbiasedness)를 만족할 경우, 수정 예측도 동일한 성질을 가진다는 장점이 있다.

최적조합을 위한 선형회귀모형은 다음과 같다.

$$\widehat{\boldsymbol{Y}}_h = \boldsymbol{S}\beta_h + \varepsilon_h, \tag{2.6}$$

여기서, $\widehat{\boldsymbol{Y}}_h$는 모든 계층의 계열인 $\boldsymbol{Y}_t$에 대한 미래 h-시점 후 기저 예측 벡터이며,

$$\beta_h = E\left[\boldsymbol{Y}_{K,T+h} | \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T\right]$$

는 최하위 단계의 기저 예측에 대한 미지의 평균이다. ε_h는 평균이 0이고 미지의 공분산 행렬 Σ_h를 갖는 선형회귀에 대한 오차이며, 미래 h-시점 예측 오차와는 다르다.

합리적인 예측일 경우 기저 예측은 근사적으로 계층적 시계열 구조를 만족할 것이며, 따라서 오차 또한 근사적으로 계층적 구조를 만족할 것이다(즉, $\varepsilon_h \approx \boldsymbol{S}\varepsilon_{K,h}$). 여기서, $\varepsilon_{K,h}$는 최하위 단계의 예측 오차를 나타낸다. 이러한 가정하에서 β_h의 최량선형불편추정량(best linear unbiased estimator; BLUE)은 다음과 같음을 보일 수 있다 (Hyndman 등, 2011).

$$\hat{\beta}_h = \left(\boldsymbol{S}'\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\widehat{\boldsymbol{Y}}_h. \tag{2.7}$$

따라서, 식 (2.7)을 식 (2.6)에 대입하여 다음과 같은 수정 예측을 얻을 수 있다.

$$\bar{\boldsymbol{Y}}_h = \boldsymbol{S}\left(\boldsymbol{S}'\boldsymbol{S}\right)^{-1}\boldsymbol{S}\widehat{\boldsymbol{Y}}_h. \tag{2.8}$$

Table 3.1. Hierarchy of the regional traffic accident counts

Level	Number of series
Korea	1
City	16
Gender	2
City × Gender	32
Total	51

또한, 수정 예측에 대한 분산은 다음과 같이 계산된다.

$$\text{var}\left(\bar{\boldsymbol{Y}}_h\right) = \boldsymbol{S}\left(\boldsymbol{S}'\sum_{h}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}. \tag{2.9}$$

최적조합 방법의 가장 큰 장점은 수정 예측의 불편성이다. 그러나, 오차의 구조가 계층적 구조를 근사적으로 만족해야 한다는 가정을 검정하기가 어렵다는 단점이 있다.

3. 실증 분석

본 장에서는 2장에서 설명한 계층적 시계열 분석 방법인 상향식 방법과 최적조합 방법을 이용하여 교통사고 발생건수를 예측하고 그 정확성을 독립적 예측인 ARIMA 모형 및 지수평활법과 비교한다. 분석에 사용된 소프트웨어는 R의 hts 패키지 (Hyndman 등, 2016)이다.

3.1. 자료 및 예측 비교 방법

실증분석에 사용한 데이터는 국내 16개 시도별 성별 교통사고 발생건수이며, 2005년부터 2015년까지의 월별 자료이다. 이 중에서 2012년 12월까지의 자료는 모형 적합을 위하여 사용하였고 2013년부터의 자료는 예측의 정확성 비교를 위한 검증 자료로 사용하였다. 자료는 교통사고분석시스템(TAAS)을 통해서 얻을 수 있다(http://taas.koroad.or.kr/).

본 자료가 가지고 있는 계층적 구조는 Table 3.1과 같다. 즉, 최상위(Korea) 단계는 132개월 동안의 국내 교통사고 발생의 총 건수이다. 1-단계(시도별, City, 단계)는 총 건수를 16개의 국내 시도별(강원, 경기, 경북, 경남, 광주, 대구, 대전, 부산, 서울, 울산, 인천, 전북, 전남, 제주, 충북, 충남)로 분해하는 단계이며, 최하위(City × Gender) 단계는 1-단계의 시도별 사고건수를 다시 성별(남성, 여성)로 분해하였다. 이와 같은 사고건수 시계열의 계층은 상하 분류가 유일하지 않으므로 그룹 시계열 자료로 볼 수 있다. 즉, 1-단계를 시도별 대신 성별(Gender)로 분해할 수도 있다. 따라서, 계층적 시계열은 총 51개 계열이 된다.

Figure 3.1은 교통사고 발생건수 자료에 대한 시계열 그림이다. 'Total' 그림은 최상위 계층으로서 국내 총 교통사고 발생건수, 'G1'와 'G2'는 각각 시도별과 성별 교통사고 발생건수, 'Bottom'은 최하위 계층으로서 시도별 성별 교통사고 발생건수를 나타낸다.

모든 계층에서 유사하게 1년을 주기로 하는 계절성의 형태가 분명히 나타나고 있다. 대체로 1월과 2월에는 상대적으로 사고건수가 적다. 남성이 전체 운전자에서 차지하는 비율이 높은 만큼 남성의 사고건수가 여성보다 훨씬 높게 나타나고 있으며, 지역상으로는 경기 및 서울에서 사고건수가 가장 높게 나타난다. 많은 계열들이 중복되어 그림에서는 잘 나타나지는 않지만, 제주도의 사고건수가 가장 낮게 나타났으며 계절성 및 남성, 여성 사고건수의 형태는 다른 지역과 비슷한 패턴을 보였다.

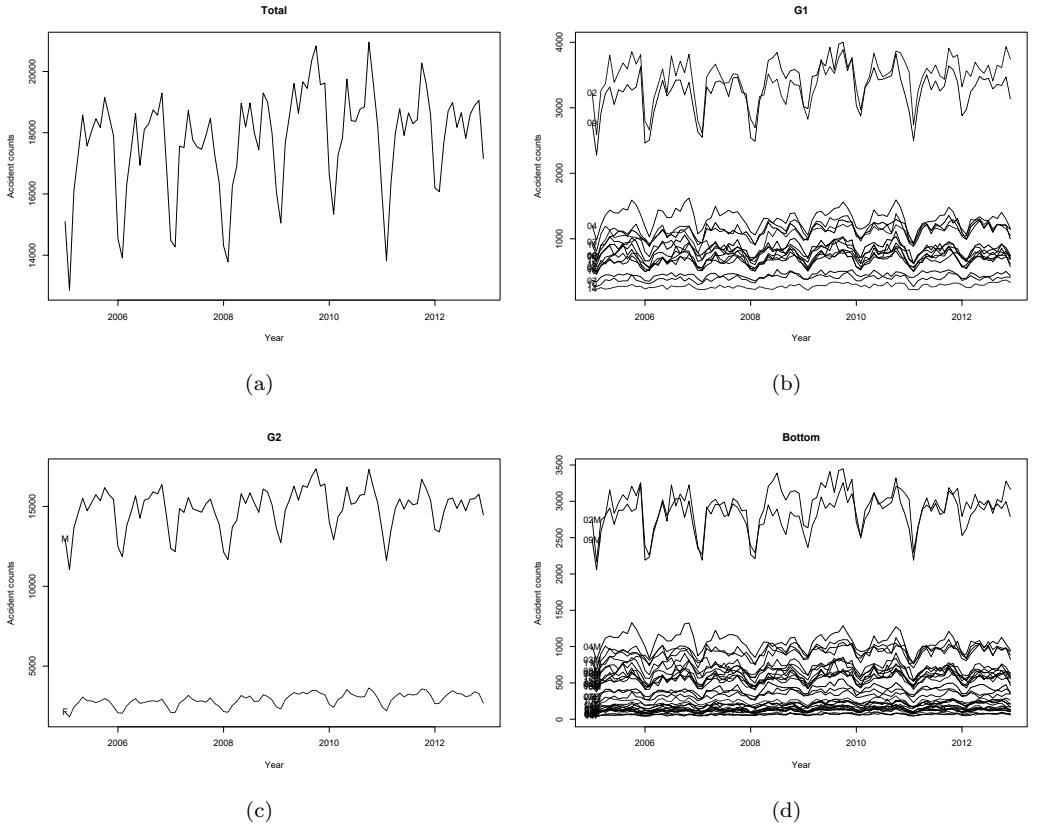

Figure 3.1. Plot of series at each level of hierarchy. Note: The numbers 01, 02, ... , 16 indicate the names of region Gangwon, Gyeonggi, Gyeongbuk, Gyeongnam, Gwangju, Daegu, Daejeon, Busan, Seoul, Ulsan, Incheon, Jeonbuk, Jeonnam, Jeju, Chungbuk, Chungnam in order, and the characters M and F indicate the genders male and female, respectively.

기저 예측을 위하여 ARIMA 모형 및 지수평활법을 각각 사용하였으며, 모형 적합을 위하여 R의 forecast 패키지 (Hyndman, 2016)의 auto.arima 및 ets 함수를 사용하였다. 이 함수들은 최적의 모형을 찾기 위하여 Akaike information criterion(AIC) 정보량 기준을 사용한다. 또한, 2013년 1월부터의 검증 자료는 rolling window 방법을 사용하여 재추정(re-estimate)과 예측을 반복하였다. 적합 모형으로부터 미래 12-시점 예측을 하였고, 이 과정을 2014년 12월까지 표본 크기를 한 달씩 증가시키면서 25번 반복하였다. 2015년 자료는 예측 과정에서의 비교를 위해서만 사용하였다. 예를 들면, 2012년 12월까지의 시계열 적합으로 2013년 1월부터 12월까지를 예측하였고, 2013년 1월까지의 시계열 적합으로 2013년 2월부터 2014년 1월까지를, 2013년 2월까지의 적합으로 2013년 3월부터 2014년 2월까지를 예측하는 것이다. 마지막 25번째 반복에서는, 2014년 12월까지의 적합으로 2015년을 예측한다. 이러한 반복 과정을 통하여 표본외 예측 정확도를 평가하였다.

3.2. 예측력 비교: MAPE

시계열 예측에 대한 정확성 비교를 위해 mean absolute percentage error(MAPE)를 사용하였다.

MAPE는 시계열 예측을 크기나 단위에 관계없이 퍼센트 오차라는 동일한 기준으로 비교 가능하다는 장점이 있다. 각 계열 $j(=1,\ldots,51)$의 미래 h-시점에 대한 $\mathrm{MAPE}_{j,h}$는 반복 예측의 평균으로 다음과 같이 정의한다.

$$\mathrm{MAPE}_{j,h} = \frac{1}{\mathrm{ITER}} \sum_{t=T}^{T+\mathrm{ITER}} \left| \frac{Y_{t+h,j} - \widehat{Y}_{t+h,j}}{Y_{t+h,j}} \right| \times 100, \quad h = 1,\ldots,12,\ \mathrm{ITER} = 25, \tag{3.1}$$

여기서 ITER = 25는 rolling window 방법의 반복 횟수를 나타내며, T는 2012년 12월부터 시작하여 한 달씩 증가된 시점으로 2014년 12월까지를 나타낸다. 따라서, 미래 h-시점 예측은 $(T+h)$-시점의 시계열에 대한 예측을 의미한다. 각 계층별 MAPE는 동일 계층 내 존재하는 계열들에 대한 MAPE의 평균으로 다음과 같이 계산하였다.

$$\mathrm{MAPE}_h(k) = \frac{1}{m_k} \sum_{i=1}^{m_k} \mathrm{MAPE}_{k:\,i,h}, \quad k = 0,1,\ldots,K(=3). \tag{3.2}$$

단, k는 각 계층을 나타내는 첨자이며 $k=0$는 최상위 계층을 나타낸다. 또한, 전체적인 예측 정확도 평가를 위하여 각 방법별 모든 계층의 MAPE 평균 및 각 계층별 모든 미래 시점의 MAPE 평균을 다음과 같이 각각 계산하였다.

$$\mathrm{MAPE}_h(\cdot) = \frac{1}{51} \sum_{i=1}^{51} \mathrm{MAPE}_{i,h}, \qquad \mathrm{MAPE}_{\cdot}(k) = \frac{1}{12} \sum_{h=1}^{12} \mathrm{MAPE}_h(k). \tag{3.3}$$

Table 3.2는 기저 예측을 위하여 ARIMA 모형을 적합한 경우, 각 단계에서 계산된 기저 예측(base)과 계층적 시계열 예측(상향식, bottom-up; 최적조합, combination)의 MAPE를 나타낸다. 비교를 용이성을 위하여 각 예측 기간에서 가장 작은 MAPE 값을 가지는 경우는 밑줄을 긋고 진하게 처리하였다. 마지막 열(mean)은 전체 예측 기간에 대한 MAPE를 평균한 값이다. 또한, Table 3.2의 맨아래쪽 행(average across all levels)은 모든 계층에 대한 세 가지 방법의 MAPE 평균값이다.

Table 3.2의 결과를 통해서 볼 때, 최상위 단계와 성별 단계에서는 독립적인 기저 예측이 상대적으로 계층적 시계열 예측보다 더 좋은 결과를 내고 있음을 확인할 수 있다. 그러나, 시도별 단계 및 시도별 × 성별(최하위) 단계에서는 최적조합 방법의 계층적 시계열 예측 방법이 절대적으로 좋은 결과를 보인다. 더 많은 계열을 포함하고 있는 단계일수록 최적조합 방법이 훨씬 더 좋은 결과를 보여주고 있다고 볼 수 있다. 모든 계층에 대한 MAPE 평균을 통해서도 최적조합 방법의 계층적 시계열이 우수함을 확인할 수 있다. 상향식 방법의 계층적 시계열 예측은 시도별 단계와 최하위 단계를 제외하고 모든 계층에서 가장 좋지 않은 예측 정확성을 보였다. 다만, 시도별 단계에서는 단기(6개월 이내) 예측에서 독립적인 기저 예측보다 좋은 정확성을 보였다. 최하위 단계에서 상향식 방법과 기저 예측이 같은 MAPE를 가지는 것은 상향식 방법의 특징이다.

Table 3.3은 기저 예측을 위하여 ARIMA 모형 대신 지수평활법을 사용하여 예측의 MAPE를 계산한 것이다. 이 경우 최상위 및 성별 계층은 ARIMA 모형에서처럼 독립적 예측이 계층적 예측보다 더 좋은 정확성을 보였다. 그러나, 모든 계층을 평균적으로 볼 때 여전히 계층적 예측이 더 우수하였다. 시도별 계층 예측에서는 상향식 방법이 절대적으로 우수한 예측력을 보였다. ARIMA 모형에서처럼, 시도별 × 성별(최하위) 단계에서 기저 예측과 상향식 예측이 동일한 MAPE를 보였다. 각 예측 방법 간 MAPE의 차이가 작다는 점은 ARIMA 모형과의 차이점으로 관찰된다.

Figure 3.2는 Tables 3.2와 3.3의 결과를 한꺼번에 그래프화한 것이며 이를 통하여 기저 예측 사이(ARIMA 모형 vs 지수평활법)의 비교를 용이하게 하였다. 결론적으로, 계층적 시계열 자료의 예

Table 3.2. MAPE for out-of-sample forecasts of the hierarchical time-series methods using ARIMA model applied to the regional traffic accident counts

	Forecast horizon (h) given at time T												
	1	2	3	4	5	6	7	8	9	10	11	12	Mean
Top level: Korea (1 series)													
Base	**2.83**	**2.47**	2.74	**2.83**	3.19	3.34	3.34	3.58	**3.48**	**3.76**	**3.87**	**4.21**	**3.30**
Bottom–up	3.61	3.36	3.03	3.32	3.26	3.48	3.52	3.68	4.05	4.04	4.33	4.44	3.68
Combination	2.94	2.75	**2.63**	2.89	**3.05**	**3.22**	**3.18**	**3.52**	3.68	3.85	4.06	4.21	3.33
Level 1: City (16 series)													
Base	5.73	6.17	6.08	6.33	6.12	6.31	6.34	6.41	6.45	6.46	6.58	6.77	6.31
Bottom–up	5.78	6.09	6.05	6.21	6.05	6.27	6.37	6.55	6.60	6.51	6.76	6.99	6.35
Combination	**5.43**	**5.69**	**5.71**	**5.90**	**5.80**	**5.99**	**6.03**	**6.23**	**6.20**	**6.23**	**6.39**	**6.58**	**6.02**
Level 2: Gender (2 series)													
Base	3.35	3.16	**3.38**	**3.71**	**3.70**	**3.99**	**4.18**	**4.45**	4.66	**4.97**	5.25	5.55	**4.20**
Bottom–up	3.88	4.00	3.91	4.09	4.10	4.27	4.45	4.48	4.84	5.00	5.23	5.60	4.49
Combination	**3.33**	**3.05**	3.41	3.86	3.88	4.12	4.21	4.46	**4.62**	4.97	**5.18**	**5.50**	4.22
Bottom level: City × Gender (32 series)													
Base	8.03	8.26	8.25	8.34	8.28	8.28	8.37	8.44	8.55	8.62	8.78	9.26	8.45
Bottom–up	8.03	8.26	8.25	8.34	8.28	8.28	8.37	8.44	8.55	8.62	8.78	9.26	8.45
Combination	**7.60**	**7.77**	**8.03**	**8.21**	**8.22**	**8.22**	**8.31**	**8.32**	**8.37**	**8.48**	**8.57**	**8.93**	**8.25**
Average across all levels (51 series)													
Base	4.99	5.02	5.11	5.30	5.32	5.48	5.56	5.72	5.78	5.95	6.12	6.45	5.57
Bottom–up	5.33	5.43	5.31	5.49	5.42	5.57	5.68	5.79	6.01	6.04	6.27	6.57	5.74
Combination	**4.83**	**4.82**	**4.95**	**5.21**	**5.24**	**5.39**	**5.43**	**5.63**	**5.72**	**5.88**	**6.05**	**6.30**	**5.45**

Table 3.3. MAPE for out-of-sample forecasts of the hierarchical time-series methods using ETS model applied to the regional traffic accident counts

	Forecast horizon (h) given at time T												
	1	2	3	4	5	6	7	8	9	10	11	12	Mean
Top level: Korea (1 series)													
Base	2.43	**1.92**	**2.19**	2.71	**2.46**	3.11	**3.09**	3.43	**3.79**	**3.70**	**4.20**	**4.23**	**3.11**
Bottom-up	**2.28**	2.05	2.29	2.66	2.69	3.09	3.23	3.46	3.88	3.87	4.28	4.42	3.18
Combination	2.40	1.96	2.24	**2.65**	2.54	**3.08**	3.14	**3.40**	3.82	3.75	4.23	4.28	3.12
Level 1: City (16 series)													
Base	4.84	5.21	5.28	5.43	5.61	5.85	6.00	5.89	6.07	6.15	6.37	6.38	5.76
Bottom-up	**4.77**	**5.01**	**5.10**	**5.35**	**5.48**	**5.71**	**5.77**	**5.75**	**6.02**	**6.04**	**6.23**	**6.22**	**5.62**
Combination	4.79	5.06	5.16	5.39	5.52	5.77	5.88	5.83	6.02	6.07	6.27	6.28	5.67
Level 2: Gender (2 series)													
Base	2.96	**2.54**	3.00	3.55	**3.46**	3.91	**4.28**	**4.45**	4.94	**5.26**	**5.63**	**5.99**	**4.16**
Bottom-up	**2.82**	2.83	3.11	**3.53**	3.67	3.93	4.41	4.72	5.06	5.47	5.86	6.22	4.30
Combination	2.89	2.58	**2.98**	3.54	3.50	**3.88**	4.29	4.51	**4.93**	5.30	5.69	6.07	4.18
Bottom level: City × Gender (32 series)													
Base	7.02	**7.03**	**7.19**	**7.31**	**7.40**	**7.56**	**7.79**	**7.66**	8.06	8.30	8.46	**8.61**	**7.70**
Bottom-up	7.02	**7.03**	**7.19**	**7.31**	**7.40**	**7.56**	**7.79**	**7.66**	8.06	8.30	8.46	**8.61**	**7.70**
Combination	**7.01**	7.05	7.28	7.40	7.43	7.59	7.85	7.69	**8.02**	**8.29**	**8.46**	8.62	7.72
Average across all levels (51 series)													
Base	4.31	4.18	**4.42**	4.75	**4.73**	5.11	5.29	**5.36**	5.71	**5.85**	6.17	**6.30**	5.18
Bottom-up	**4.22**	4.23	4.42	**4.71**	4.81	**5.07**	5.30	5.40	5.75	5.92	6.20	6.37	5.20
Combination	4.27	**4.16**	4.42	4.74	4.75	5.08	**5.29**	5.36	**5.70**	5.85	**6.16**	6.31	**5.17**

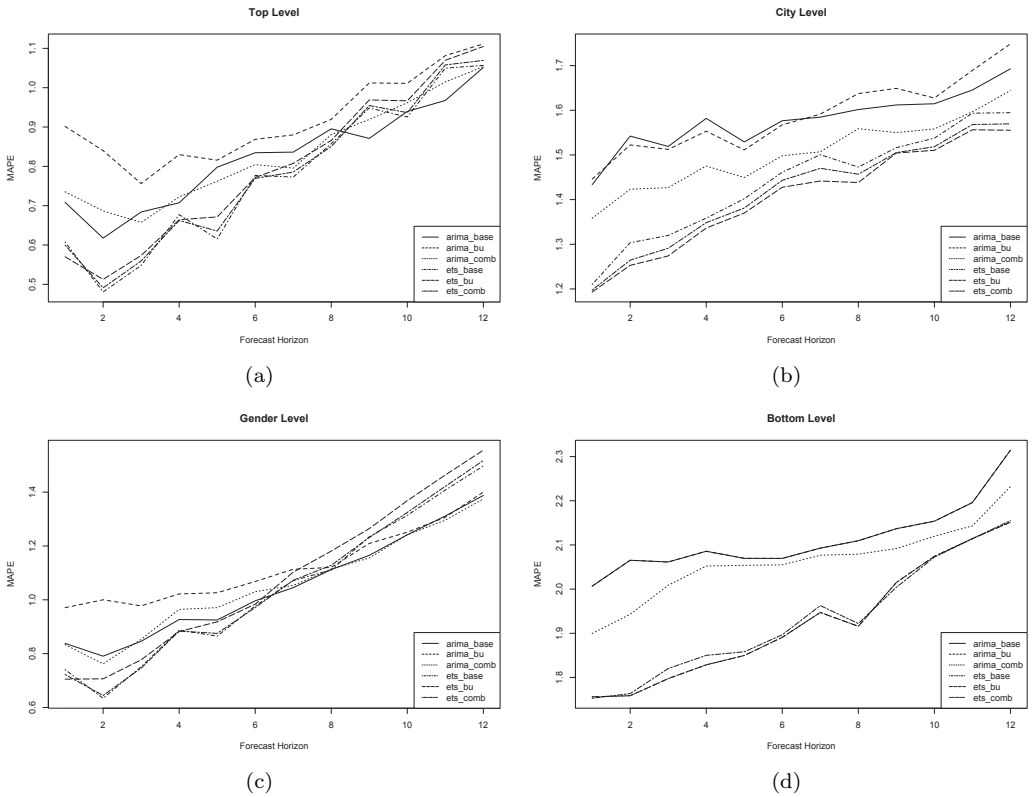

Figure 3.2. Plots of MAPE for all hierarchical levels using ARIMA and ETS models.

측에서 지수평활법을 이용할 때가 ARIMA 모형을 이용할 때보다 전반적으로 우수한 예측력을 보였다. 그러나, 장단기 예측 및 각 계층을 구분해서 살펴볼 때 몇 가지 흥미로운 발견은 (1) 최상위 계층의 장기 예측에서 독립적인 ARIMA 모형이 우수하였고; (2) 성별 계층에서는 단기 예측에서 지수평활법이, 장기 예측에서는 ARIMA 모형이 각각의 예측 방법에서 우수하였다는 점이다. 상대적으로 볼 때 지수평활법이 단기 예측에, ARIMA 모형이 장기 예측에서 더 유리한 경향을 보인다.

3.3. 예측력 비교: 통계적 가설 검정

예측 방법 간 예측 정확성 차이의 유의성을 검정하기 위해 Friedman (1937)의 검정을 수행하였다. Friedman의 검정은 비모수적 랜덤화 블록 설계로서 일원배치 분산분석의 비모수적 방법이다. 검정 통계량은 3가지 예측 방법에 의한 Table 3.2와 3.3의 MAPE를 각 미래 예측 시점에서 가장 작은 값을 1순위로, 가장 큰 값을 3순위로 매긴 후 각 예측 방법별 순위합을 이용하여 구한다. MAPE가 같을 경우 평균 순위로 매긴다. 따라서, 예측 방법 간 정확성 차이가 있다면 적어도 한 방법의 순위합의 제곱은 유의하게 작은 값을 가질 것이다. Friedman의 검정 통계량은 다음과 같이 표현할 수 있다.

$$F = \frac{12}{HK(K+1)} \sum_{j=1}^{K} R_j^2 - 3H(K+1).$$

Table 3.4. Friedman test statistics to test statistical significance of the forecast accuracy difference

	ARIMA model		ETS model	
	$\sum_j R_j^2$	F	$\sum_j R_j^2$	F
Top level	1946	18.17	1826	8.17
City level	1946	18.17	2016	24.00
Gender level	1910	15.17	1826	8.17
Bottom level	1944	18.00	1752	2.00

단, $K = 3$은 고려한 예측 방법의 수이며, $H = 12$는 미래 예측 기간의 수이다. R_j는 각 예측 방법의 MAPE 순위합이다. Friedman 검정 통계량은 "예측 방법에 유의한 차이가 없다"라는 귀무가설이 참일 때, 근사적으로 자유도가 $K-1$인 χ^2_{K-1} 분포를 따른다. Table 3.4는 각 계층에서 Friedman 검정을 실시한 결과이다.

따라서, 유의수준 5%(기각치 = 5.99)에서 기저 예측 방법(ARIMA 모형 및 지수평활법)에 관계없이 모든 계층에서 예측 방법 간 유의한 차이가 있다는 것을 확인할 수가 있다. 단, 지수평활법을 사용할 경우 최하위 계층에서는 예측 방법 간에 유의한 차이를 없는 것으로 나타난다. 참고로, 이 절차는 R에서 stats 패키지의 Friedman.test 함수를 이용하여 수행하였다.

세 가지 예측 방법 중 어느 방법이 유의한 차이를 보이는지 파악하기 위해 사후 쌍별 검정인 Nemenyi (1963) 검정을 수행하였다. 이것은 보통의 분산분석에서 사후 검정(post-hoc test)에 해당하는 것으로서, 이 결과를 이용하여 예측 방법간 정확성의 순서를 찾아낼 수 있다. 정확성의 순서는 Table 3.2와 3.3의 마지막 열에 계산된 MSPE 값의 크기 순서를 이용하였다. Nemenyi 검정은 "어떤 두 가지 방법이 비슷한 예측 정확성을 갖는다"는 귀무가설을 가지는 양측 검정이다. 두 방법의 예측 정확성은 해당 평균 순위가 최소한 기각 차이만큼의 차가 있을 경우 유의한 차이가 존재한다고 판단을 한다. 검정 통계량은 다음과 같으며, 기각값 q_α은 Demsar (2006)이 제안한 값을 사용하였다.

$$q_\alpha \sqrt{\frac{K(K+1)}{6H}}.$$

Table 3.5와 3.6은 각 기저 예측 방법(ARIMA 모형 및 ETS 모형)에 따른 Nemenyi 검정에 의한 p-값이며, 이것은 R에서 PMCMR 패키지 (Pohlert, 2016)의 posthoc.friedman.nemenyi.test 함수를 사용한 결과이다.

Table 3.5를 통하여 다음과 같은 결론을 내릴 수 있다; (1) 계층적 예측시 기저예측으로 ARIMA 모형을 이용할 경우, 상향식 방법보다는 최적조합 방법이 우수하다; (2) 최상위 계층 및 성별 계측에서는 독립적인 예측이 계층적 예측 방법보다는 우수하다.

Table 3.6을 통해서는 다음과 같은 결론을 내릴 수 있다; (1) 계층적 예측시 기저예측으로 지수평활법을 이용할 경우, ARIMA 기저 예측과는 달리 상향식 방법과 최적 조합의 예측 정확성 차이는 크지 않다; (2) 최상위 및 성별 계층에서 독립적인 예측은 상향식 방법보다는 우수한 결과를 보였다. 단, 독립적인 예측과 최적조합 예측 방법 간 예측력에 있어서 유의한 차이를 보이지 않았다.

4. 결론

최근 데이터 규모가 커짐과 동시에 다양한 형태의 시계열 자료가 나타나고 있으며 이에 대한 분석과 예측의 수요가 높아지고 있다. 본 논문에서는 계층적 또는 그룹화된 시계열 자료의 예측 방법을 소개하고

Table 3.5. *p*-values of the Nemenyi's test statistics to test statistical significance of the forecast accuracy among methods using ARIMA model

		Base	Bottom–up
Top level: Korea	Bottom–up	**0.0003**	-
	Optimal combination	0.9122	**0.0015**
Level 1: City	Bottom–up	0.9122	-
	Optimal combination	**0.0015**	**0.0003**
Level 2: Gender	Bottom–up	**0.0015**	-
	Optimal combination	0.9773	**0.0031**
Bottom level: City × Gender	Bottom–up	1.0000	-
	Optimal combination	**0.0001**	**0.0001**

Table 3.6. *p*-values of the Nemenyi's test statistics to test statistical significance of the forecast accuracy among methods using ETS model

		Base	Bottom-up
Top level: Korea	Bottom–up	**0.0220**	-
	Optimal combination	0.9120	0.0640
Level 1: City	Bottom–up	**<0.0001**	-
	Optimal combination	**0.0380**	**0.0380**
Level 2: Gender	Bottom–up	**0.0220**	-
	Optimal combination	0.9120	0.0640
Bottom level: City × Gender	Bottom–up	0.6900	-
	Optimal combination	0.2300	0.6900

있으며 이의 실증 분석으로 국내 16개 시도별 및 성별 교통사고 발생건수 예측에 적용하였다. 실증 분석을 통하여 계층적 시계열 예측 및 독립적 시계열 예측 방법을 서로 비교하였으며, 이를 통하여 계층적 시계열 예측의 우수성과 특징을 살펴보았다. 더불어 계층적 예측을 위한 기저 예측으로서 ARIMA 모형과 지수평활법의 특징도 살펴보았다. 그러나 기저 예측이 두 가지 예측에만 제한되어 있는 만큼 향후 다양한 기저 예측 방법의 개발이 절실하며, R 이외의 다른 소프트웨어에서 사용할 수 있는 계층적 시계열 예측을 위한 프로그램 개발 및 보급이 시급하다. 계층적 또는 그룹화된 시계열 자료 분석에서 향후 잠재력있는 연구 주제로서는, 계층 및 그룹의 축소화와 이에 대한 검정 방법의 개발을 들 수 있다. 이는 통계학에서 모수 절약의 원칙 관점에서 의미가 있다고 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.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets, *Journal of Machine Learning Research*, **7**, 1–30.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association*, **32**, 675–701.

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

Han, S. J. and Kim, K. J. (2007). Road accident trends analysis with time series models for various road types, *International Journal of Highway Engineering*, **9**, 1–12.

Hyndman, R. J. (2016). Forecast: Forecasting functions for time series and linear models, *R package version 7.3*, Available from: https://CRAN.R-project.org/package=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. and Athanasopoulos, G. (2014). *Forecasting Principles and Practice*, Otexts, Available from: https://www.otexts.org/fpp

Hyndman, R. J, Wang, E., Lee, W., and Wichramasuriya, S. (2016). hts: Hierarchical and grouped time series, *R package version 5.0*, Available from: https://CRAN.R-project.org/package=hts

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

Nemenyi, P. B. (1963). *Distribution-free Multiple Comparisons* (PhD thesis), Princeton University, New Jersey.

Pohlert, T. (2016). PMCMR: Calculate pairwise multiple comparisons of mean rank sums, *R package version 4.1*, Available from: https://CRAN.R-project.org/package=PMCMR

Shang, H. L and Smith, P. W. F. (2013). Grouped time-series forecasting with an application to regional infant mortality counts, *Southampton, GB, ESRC Centre for Population Change 40*.

Weale, M. (1988). The reconciliation of values, volumes and prices in the national accounts, *Journal of the Royal Statistical Society, Series A*, **151**, 211–221.

Zellner, A. and Tobias, J. (2000). A note on aggregation, disaggregation and forecasting performance, *Journal of Forecasting*, **19**, 457–469.

계층적 시계열 분석을 이용한 지역별 교통사고 발생건수 예측

이주은[a] · 성병찬[a,1]

[a]중앙대학교 응용통계학과

(2017년 1월 5일 접수, 2017년 1월 16일 수정, 2017년 1월 23일 채택)

요 약

본 논문에서는 계층적 시계열 자료 분석을 위한 대표적인 두 가지 방법인 상향식과 최적조합 예측법을 소개한다. 이러한 예측법은 계층적 시계열을 구성하는 모든 계열을 예측해야 하는 독립적 예측과 달리, 임의의 조정 과정이 없이 하위 계층 계열의 예측값의 합은 항상 상위 계층의 예측값과 일치하게 된다. 또한, 독립적 예측과 비교하여 예측력을 향상시킨다. 계층적 예측법의 효율성을 살펴보기 위하여 국내 16개 시도별 남녀 교통사고 발생건수 시계열 자료를 예측하였다. 이를 통하여 교통사고 발생건수에 대한 각 계층의 예측에서 계층적 방법과 독립적 방법의 차이점 및 우수성을 비교하였다.

주요용어: 그룹화 시계열, 수정예측, 상향식 예측, 최적조합 예측, ARIMA 모형, 지수평활법

이 논문은 2015년도 중앙대학교 CAU GRS 지원에 의하여 작성되었음.

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