The Korean Journal of Applied Statistics
2024, Vol. 37, No. 4, 499–508

DOI: https://doi.org/10.5351/KJAS.2024.37.4.499
Print ISSN1225-066X / Online ISSN 2383-5818

Forecasting hierarchical time series for foodborne disease outbreaks

In-Kwon Yeo[1,a]

[a]Department of Statistics, Sookmyung Women's University

Abstract

In this paper, we investigate hierarchical time series forecasting that adhere to a hierarchical structure when deriving predicted values by analyzing segmented data as well as aggregated datasets. The occurrences of food poisoning by a specific pathogen are analyzed using zero-inflated Poisson regression models and negative binomial regression models. The occurrences of major, miscellaneous, and overall food poisoning are analyzed using Poisson regression models and negative binomial regression models. For hierarchical time series forecasting, the MinT estimation proposed by Wickramasuriya *et al.* (2019) is employed. Negative predicted values resulting from hierarchical adjustments are adjusted to zero, and weights are multiplied to the remaining lowest-level variables to satisfy the hierarchical structure. Empirical analysis revealed that there is little difference between hierarchical and non-hierarchical adjustments in predictions based on pathogens. However, hierarchical adjustments generally yield superior results for predictions concerning major, miscellaneous, and overall occurrences. Without hierarchical adjustment, instances may occur where the predicted frequencies of the lowest-level variables exceed that of major or miscellaneous occurrences. However, the proposed method enables the acquisition of predictions that adhere to the hierarchical structure.

Keywords: hierarchical time series forecasts, negative binomial regression, optimal combination, Poisson regression, zero-inflated regression

1. 서론

지구온난화로 인한 기후변화는 유래없는 자연재해와 생태계의 변화를 초래하고 농수산물의 재배와 포획, 채취 지역의 변화를 가져왔다. 보건 분야에서는 기후에 영향을 받는 식품 위해 발생을 체계적으로 분석하고 대응전략 수립하기 위한 연구가 필요했다. 이와 관련하여 식품의약품안전평가원의 지원 하에 한국보건산업진흥원이 수행한 '기후변환에 따른 식중독 발생 영향 분석 및 관리 체계 연구'에서는 기후변화와 식중독 발생에 관련된 종합적 분석 결과가 제시되었는데 자세한 내용은 Jeong과 Oh (2009)의 연구보고서를 참고하기 바란다.

많은 식중독 발생 연구에서 식중독과 기후 간 관련이 있는 것으로 분석되었다. Fleury 등 (2006)은 포아송 분포와 로그연결함수를 가정한 일반화선형모형(generalized linear model)과 일반화가법모형(generalized additive model)를 이용하여 기온과 계절성이 여러 원인균별 식중독 발생건수에 어떻게 영향을 주는지를 분석하였다. Zhang 등 (2007)은 남호주의 주간 살모넬라 로그발생건수를 52주 SARIMA(4,0,0)(1,0,0)으로 분석하여 2주전 최고온도와 계절성이 유의하게 영향을 주는 것을 확인하였다. Zhang 등 (2008)은 살모넬라균에

[1]Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-Gu, Seoul 04310, Korea.
E-mail: inkwon@sookmyung.ac.kr

Published 31 August 2024 / journal homepage: http://kjas.or.kr

의한 전염과 기후변화 간 관계를 포아송회귀모형, 자기회귀 포아송회귀모형, 다중선형회귀모형, SARIMA 모형으로 분석하여 비교하였다. Choi 등 (2008)은 일별 식중독 발생과 기상자료를 로그선형모형으로 분석하여 식중독 발생의 위협을 예보할 수 있는 사고발생지수를 개발하였다. Yeo (2012)는 발생빈도에 대해 AR(4) 모형, 로그발생빈도에 대해 AR(4) 모형, 평균기온, 습도, 일조량과 월을 가변수로 설명변수에 포함시킨 포아송회귀모형과 AR(4)-X 모형으로 주간 식중독 발생 건수를 분석하였다.

식중독은 기후와 같은 자연적 요소와 더불어 조리법이나 섭취형태 등의 식문화와 관련이 있으며 Jeong과 Oh (2009)에 따르면 국가별로 식중독 원인에 차이가 있는 것으로 조사되었다. 특히 생식을 좋아하는 우리나라의 경우 세균이나 바이러스로 인한 식중독 사례가 많은데 국내에서 자주 발생하는 식중독의 원인균으로 노로바이러스(norovirus), 병원성 대장균(Pathogenic Escherichia coli), 살모넬라(Salmonella), 장염비브리오(Vibrio parahaemolyticus)와 황색포도상구균(Staphylococcus aureus)이 있다. 식중독 관련 주요 연구에서도 이들 원인균으로 인한 식중독을 우선적으로 다루고 있다. 본 연구에서는 이들 5가지 원인균에 의해 발생한 식중독을 주요 식중독으로 분류하고 나머지 캠필로박터제주니(Campylobacter Jejuni), 클로스트리디움퍼프린젠스(Clostridum perfringens), 바실러스세레우스(Bacillus cereus), 기타 세균, 기타 바이러스, 원충(protozoa), 자연독(natural toxin), 화학물질 그리고 불명으로 이루어진 9가지 원인물질에 인한 식중독을 기타 식중독으로 나누어 분석하고자 한다.

원인물질별로 나누어 식중독을 분석한 국내 연구로는 Jeong과 Oh (2009)와 Jung 등 (2012), Yeo (2013)가 있다. Jeong과 Oh (2009)는 해당 월과 전 월의 평균기온과 습도, sine과 cosine 함수를 이용한 계절성, 연도 등을 설명변수로 설정한 포아송 회귀모형과 평균기온, 최고온도, 최저온도, 강수량을 설명변수로 추가한 SARIMA(0, 1, 1)(0, 1, 1)-X 모형을 기반으로 노로바이러스, 병원성 대장균, 살모넬라, 장염비브리오, 황색포도상구균, 기타 및 불명으로 인한 발생건수를 추정하였다. Jung 등 (2012)은 원인물질별로 분석할 때 포아송이나 음이항 회귀모형에서 설명할 수 있는 것 보다 상대적으로 많은 0 관측값이 발생하는 것을 확인하였으며 노로바이러스와 살모넬라군에 의한 발생건수를 영과잉 포아송과 음이항 회귀모형으로 분석하였다.

식중독 발생건수를 원인물질별로 분석할 수도 있고 이를 합한 전체 발생건수를 분석할 수도 있다. 위에서 언급한 연구에서는 원인물질별 발생건수의 분석과 전체 발생건수의 분석이 별개로 수행되어 개별 원인에 의한 식중독 적합건수 합이 전체 발생건수를 분석해서 나온 적합건수와 일치하지 않는 한계가 있다. 이 논문에서는 원인물질별 분석결과에 따른 적합건수 또는 예측건수의 합이 전체 발생건수를 분석하여 유도한 적합건수 또는 예측건수를 같게 만드는 방법에 대해 알아본다.

2. 제안 방법론

Y_{it}를 t 시점에서 i-번째 원인물질에 의한 식중독 발생건수라고 하고 Y_{it}의 기댓값을 $E(Y_{it}) = \mu_{it}$라고 하자. 주요 원인균($i = 1, \ldots, 5$)에 의한 식중독 발생건수를 $Y_t^{(1)}$이라고 하고 기타 원인($i = 6, \ldots, 14$)에 의한 식중독 발생건수를 $Y_t^{(2)}$라고 하면 두 확률변수의 기댓값은 다음과 같다.

$$E\left(Y_t^{(1)}\right) = \mu_t^{(1)} = \sum_{i=1}^{5} \mu_{it}, \quad E\left(Y_t^{(2)}\right) = \mu_t^{(2)} = \sum_{i=6}^{14} \mu_{it} \tag{2.1}$$

또한 t 시점에서 전체 식중독 발생건수를 $Y_t^{(0)}$라고 하면 다음의 조건을 만족해야 한다.

$$E\left(Y_t^{(0)}\right) = \mu_t^{(0)} = \mu_t^{(1)} + \mu_t^{(2)} = \sum_{i=1}^{14} \mu_{it}, \tag{2.2}$$

여기서 각 기댓값은 설명변수 $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})^T$에 영향을 받고 임의의 연결함수 $g(\cdot)$를 통해 다음과 같은 관계식을 가진다고 가정한다.

$$g(\mu_{it}) = \beta_{i0} + \beta_{i1}x_{t1} + \cdots + \beta_{ip}x_{tp} = \boldsymbol{\beta}_i^T \boldsymbol{x}_t.$$

문제는 원인물질별, 주요 원인균, 기타 원인, 전체 식중독 건수에 대한 개별 분석를 통해 유도된 $\hat{\mu}_{it}$, $\hat{\mu}_t^{(1)}$, $\hat{\mu}_t^{(2)}$, $\hat{\mu}_t^{(0)}$은 (2.1)과 (2.2)의 등식 조건을 만족할 가능성이 매우 희박하다는 것이다. 이 연구에서는 개별 분석에서 유도된 이들 적합값 및 예측값을 Wickramasuriya 등 (2019)이 제안한 최적결합(optimal combination) 방법을 이용하여 (2.1)과 (2.2)의 등식 조건을 만족하도록 조정하는 방법을 알아본다.

2.1. 원인균별 식중독 건수 추정

서론에서도 언급한 것과 같이 원인균별로 식중독 발생건수를 분석할 경우 포아송 회귀모형이나 음이항 회귀모형으로 분석에 의해 기대되는 0의 비율보다 실제 관측 비율이 높기 때문에 이들 모형의 적합력이 떨어지고 과대산포(over-dispersion)의 문제가 발생한다. 이러한 문제를 해결하기 위해 본 연구에서는 Jung 등 (2012)에서 사용했던 영과잉 포아송 회귀모형과 영과잉 음이항 회귀모형으로 원인물질별 식중독 발생건수를 분석한다. 영과잉 포아송 모형은 Lambert (1992)에 의해 제안되었으며 Agarwal 등 (2003)에 의하면, 영과잉 포아송 분포 자료는 다음과 같이 단계를 거쳐 자료가 생성되며 이를 근거로 가능도함수를 유도할 수 있다.

- 베르누이 난수생성: 0일 확률이 π인 베르누이 확률분포에서 난수를 생성하여 0이면 그 자료를 사용하고 1이면 다음 단계로 넘어 감
- 포아송 난수생성: 평균이 ν인 포아송 난수를 생성하여 자료로 사용함

만약 영과잉 음이항 분포 자료를 얻고자 한다면 두 번째 과정에서 포아송 대신 음이항 난수를 생성하여 사용한다.

두 번째 단계에서 평균이 ν인 포아송 또는 음이항 분포의 확률 질량함수를 $f^*(y;\nu)$라고 하면 영과잉 포아송 또는 음이항 분포의 확률질량함수는 다음과 같다.

$$f(y;\pi,\nu) = \begin{cases} \pi + (1-\pi)f^*(y;\nu), & y = 0, \\ (1-\pi)f^*(y;\nu), & y = 1,2,3,\ldots \end{cases}$$

영과잉 포아송 회귀모형이나 음이항 회귀모형은 Y_{it}가 위의 분포를 따르고 모수인 π_{it}와 ν_{it}가 설명변수 $\boldsymbol{x}_t$에 영향을 받으며 적절한 연결함수 $g_1(\cdot)$와 $g_2(\cdot)$를 통해 다음과 같은 관계를 가정한다.

$$g_1(\pi_{it}) = \alpha_{i0} + \alpha_{i1}x_{t1} + \cdots + \alpha_{ip}x_{tp} = \boldsymbol{\alpha}_i^T \boldsymbol{x}_t,$$
$$g_2(\nu_{it}) = \beta_{i0} + \beta_{i1}x_{t1} + \cdots + \beta_{ip}x_{tp} = \boldsymbol{\beta}_i^T \boldsymbol{x}_t,$$

여기서 $g_1(\cdot)$은 일반적으로 로짓(logit)연결함수, $g_1(\pi) = \log(\pi/(1-\pi))$, $g_2(\cdot)$는 로그연결함수, $g_2(\nu) = \log(\nu)$가 많이 사용된다. 분석에서 π_{it}와 ν_{it}에 대해 동일한 설명변수를 가정할 필요는 없으며 실증분석에서도 AIC 기반 변수선택을 통해 서로 다른 설명변수를 선택하도록 하였다. 여기에서는 설명의 편의를 위해 동일한 설명변수를 가정한다. 이 모형을 기반으로 한 자세한 모수 추정 내용은 Jung 등 (2012)을 참조하기 바란다.

회귀모수 $\boldsymbol{\alpha}_i = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{ip})^T$와 $\boldsymbol{\beta}_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})^T$의 최대가능도 추정량을 $\hat{\boldsymbol{\alpha}}_i = (\hat{\alpha}_{i0}, \hat{\alpha}_{i1}, \ldots, \hat{\alpha}_{ip})^T$와 $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{i0}, \hat{\beta}_{i1}, \ldots, \hat{\beta}_{ip})^T$라고 하면 Y_{it}의 평균 $E(Y_{it}) = \mu_{it} = (1-\pi_{it})\nu_{ij}$는 다음과 같이 추정된다.

$$\hat{\mu}_{it} = \left(1 - g_1^{-1}\left(\hat{\boldsymbol{\alpha}}_i^T \boldsymbol{x}_t\right)\right) g_2^{-1}\left(\hat{\boldsymbol{\beta}}_i^T \boldsymbol{x}_t\right).$$

2.2. 주요, 기타, 전체 식중독 건수 추정

주요, 기타 및 전체 식중독 발생 건수는 각종 원인균에 의한 식중독 발생 건수의 합으로 월별 자료의 경우 일반적인 포아송 회귀모형이나 음이항 회귀모형으로도 충분히 설명 가능한 것으로 분석되었다. $Y_t^{(j)}$를 t 시점에서의 전체($j = 0$), 주요($j = 1$), 기타($j = 2$) 식중독 발생 건수라고 하면, $E(Y_t^{(j)}) = \mu_t^{(j)}$인 포아송 또는 음이항 분포를 따르고 평균 $\mu_t^{(j)}$는 다음과 같이 연결함수 $g(\cdot)$를 통해 설명변수 $\boldsymbol{x}_t$에 영향을 받는다고 가정한다.

$$g\left(\mu_t^{(j)}\right) = \beta_0^{(j)} + \beta_1^{(j)} x_{t1} + \cdots + \beta_p^{(j)} x_{tp} = \boldsymbol{\beta}_{(j)}^T \boldsymbol{x}_t.$$

일반적으로 사용되는 연결함수 $g(\cdot)$는 로그함수이다. 음이항 분포는 모수설정에 따라 다양한 형태로 표시되는데 일반화선형모형에서는 다음과 같은 재모수화한 확률질량함수로 표시된다.

$$f(y;\mu,\phi) = \frac{\Gamma(y + 1/\phi)}{y!\Gamma(1/\phi)} \left(\frac{1}{1+\phi y}\right)^{1/\phi} \left(\frac{\phi\mu}{1+\phi\mu}\right),$$

이렇게 설정하면 $\mathrm{Var}(Y_t^{(j)}) = \mu_t^{(j)}(1 + \mu_t^{(j)}\phi)$가 되는데 $\phi = 0$이면 포아송 회귀모형과 동일한 분산이 되고 $\phi > 0$이면 포아송 회귀분석으로 설명할 수 없는 과대산포를 모형화 할 수 있다.

일반화선형모형에서의 회귀모수 추론은 기본적인 내용이기 때문에 설명은 생략한다. 회귀모수 $\boldsymbol{\beta}^{(j)} = (\beta_0^{(j)}, \beta_1^{(j)}, \ldots, \beta_p^{(j)})^T$의 최대가능도 추정량을 $\hat{\boldsymbol{\beta}}^{(j)} = (\hat{\beta}_0^{(j)}, \hat{\beta}_1^{(j)}, \ldots, \hat{\beta}_p^{(j)})^T$라고 하면 모든 시점 t에서 다음의 관계가 성립해야 한다.

$$\hat{\mu}_t^{(0)} = g^{-1}\left(\hat{\boldsymbol{\beta}}^{(0)}\boldsymbol{x}_t\right) = \hat{\mu}_t^{(1)} + \hat{\mu}_t^{(2)} = g^{-1}\left(\hat{\boldsymbol{\beta}}^{(1)}\boldsymbol{x}_t\right) + g^{-1}\left(\hat{\boldsymbol{\beta}}^{(2)}\boldsymbol{x}_t\right),$$

또한 앞서 추정한 개별 원인균별 평균발생건수 $\hat{\mu}_{it}$와 모든 시점 t에서 다음의 관계가 성립해야 하지만 이를 만족하는 경우는 거의 없다.

$$\hat{\mu}_t^{(1)} = \sum_{i=1}^{5} \hat{\mu}_{it}, \quad \hat{\mu}_t^{(2)} = \sum_{i=6}^{14} \hat{\mu}_{it}, \quad \hat{\mu}_t^{(0)} = \sum_{i=1}^{14} \hat{\mu}_{it}. \tag{2.3}$$

2.3. 계층구조 하에서의 적합값과 예측값 조정

식중독 발생건수에 대한 추정 및 예측이 고도화 되기 위해서는 원인물질별, 주요, 기타 및 전체 식중독 발생건수의 적합값들이 (2.3)의 관계를 만족해야 한다. 계층구조(hierarchical structure)를 가지는 여러 시계열 자료를 이용하여 예측값을 유도할 때 계층구조를 만족하도록 예측하는 방법을 계층 시계열 예측(forecasting hierarchical time series)라고 한다. 계층 시계열 예측에서 가장 간단한 방법은 최하위 시계열의 예측값들을 우선 유도하고 그 결과를 결합하는 bottom-up 방식이다. 이와 관련된 연구로는 Orcutt 등 (1968), Dunn 등 (1976), Shlifer와 Wolff (1979)가 있는데 이 방법은 최하위 시계열 간 존재할 수 있는 연관성을 고려하지 않고 독립적으로 분석하여 결과를 결합하는 문제가 있다. 또 다른 방법으로 최상위 시계열의 예측값을 유도하고 그 결과를 하위 시계열로 분할하는 top-down 방식이 있는데 Hyndman 등 (2011)에 의하면 분할된 예측값에 편향이 발생하는 것으로 나타났다.

이 논문에서는 Wickramasuriya 등 (2019)이 제안한 최적결합 방법을 이용하여 원인물질별, 주요, 기타 및 전체 식중독 발생건수의 적합값과 예측값을 조정한다. 개별 분석을 통해 유도한 시점 t에서의 적합값 또는 예측값 벡터를 $\hat{\boldsymbol{\mu}}_t = (\hat{\mu}_t^{(0)}, \hat{\mu}_t^{(1)}, \hat{\mu}_t^{(2)}, \hat{\mu}_{1t}, \ldots, \hat{\mu}_{14t})$라고 하고 최적결합을 통해 계층구조를 만족하도록 조정한 벡터를 $\tilde{\boldsymbol{\mu}}_t = (\tilde{\mu}_t^{(0)}, \tilde{\mu}_t^{(1)}, \tilde{\mu}_t^{(2)}, \tilde{\mu}_{1t}, \ldots, \tilde{\mu}_{14t})$라고 하자. Hyndman 등 (2011)에 의하면 이 두 벡터는 다음과 같이 설정할 수 있다.

$$\tilde{\boldsymbol{\mu}}_t = \mathbf{S}\mathbf{P}\hat{\boldsymbol{\mu}}_t,$$

Table 1: Goodness of fit

Pathogen	Poisson						Negative binomial					
	$\sqrt{\text{MSE}}$		MAE		MAX		$\sqrt{\text{MSE}}$		MAE		MAX	
	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$
PE	2.62	2.62	1.64	1.65	13.40	13.26	2.69	2.71	1.69	1.71	13.75	13.57
Salmonella	1.68	1.67	1.16	1.16	8.41	8.30	1.71	1.69	1.17	1.17	8.67	8.55
VP	1.65	1.64	0.74	0.74	14.91	14.91	1.99	1.99	0.76	0.77	17.98	17.90
Norovirus	3.36	3.42	2.22	2.31	19.30	19.67	3.42	3.42	2.23	2.28	19.48	19.92
SA	1.19	1.19	0.85	0.86	6.24	6.22	1.42	1.43	0.89	0.90	12.53	12.66
CJ	1.12	1.12	0.73	0.73	7.41	7.41	1.11	1.11	0.73	0.73	7.07	7.10
CP	1.13	1.14	0.78	0.79	5.81	5.81	1.16	1.16	0.76	0.76	6.90	6.77
BC	0.74	0.74	0.54	0.54	3.86	3.85	0.74	0.74	0.54	0.54	3.86	3.86
OB	0.40	0.40	0.21	0.21	2.01	2.00	0.40	0.40	0.21	0.22	2.02	2.01
OV	0.48	0.48	0.32	0.32	2.25	2.25	0.50	0.50	0.35	0.35	2.73	2.73
Protozoa	1.07	1.07	0.48	0.49	8.81	8.74	1.08	1.08	0.49	0.49	8.82	8.75
NT	0.42	0.42	0.28	0.28	2.67	2.67	0.41	0.41	0.26	0.26	2.71	2.71
CS	0.13	0.13	0.03	0.03	1.00	1.00	0.13	0.13	0.03	0.03	1.00	1.00
Unknown	4.71	4.81	3.74	3.82	19.86	19.92	4.80	4.91	3.80	3.89	19.90	20.01
Major	6.25	5.95	4.59	4.30	27.31	26.61	6.34	6.07	4.69	4.42	27.67	26.28
Miscellaneous	5.74	5.48	4.62	4.47	19.63	19.17	5.95	5.57	4.70	4.51	19.55	19.15
Overall	9.66	9.32	7.40	7.13	38.18	36.33	9.94	9.44	7.54	7.26	39.35	36.51

PE: Pathogenic Escherichia coli, VP: Vibrio parahaemolyticus, SA: Staphylococcus aureus,
CJ: Campylobacter Jejuni, CP: Clostridum perfringens, BC: Bacillus cereus,
OB: Other bacteria, OV: Other virus, NT: natural toxin, CS: chemical substance.

여기서 행렬 $\mathbf{P}$는 $\hat{\boldsymbol{\mu}}_t$를 최하위 계층의 적합 및 예측값으로 변환해주는 역할을 하고 $\mathbf{S}$는 행렬 $\mathbf{P}$에 의하여 유도된 적합 및 예측값이 계층구조를 만족하도록 만들어 주는 역할을 한다. 현재 분석에서는 14개 원인물질을 5개의 주요 원인균과 9개 기타 원인물질로 나누어 계층구조를 구성하기 때문에 $\mathbf{S}$는 다음과 같은 17×14 행렬로 표시할 수 있다.

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & \dots & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & \dots & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

최적결합은 예측 오차의 변동성을 최소화시키는 $\mathbf{P}$를 찾는 것으로 Wickramasuriya 등 (2019)은 오차 공분산 행렬의 대각합(trace)을 최소화하는 방법을 제안하였으며 이를 통해 다음과 같은 추정값을 얻을 수 있다.

$$\tilde{\boldsymbol{\mu}}_t = \mathbf{S}\left(\mathbf{S}^T\mathbf{W}_t^{-1}\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{W}_t^{-1}\hat{\boldsymbol{\mu}}_t, \tag{2.4}$$

여기서 $\mathbf{W}_t$는 t 시점에서의 예측오차 공분산으로 보통최소제곱(ordinary least squares)이나 가중최소제곱(weigh ted least squares) 방법으로 추정한다. 이렇게 유도된 추정량을 'MinT'(minimum trace) 추정량이라고 한다. 이에 대한 자세한 내용과 추정량의 성질은 Wickramasuriya 등 (2019)을 참조하기 바란다. Kim 등 (2022)은 이 MinT 추정을 통해 가계동향조사의 지출부분에 대한 계층 시계열 예측 연구를 수행하기도 하였다.

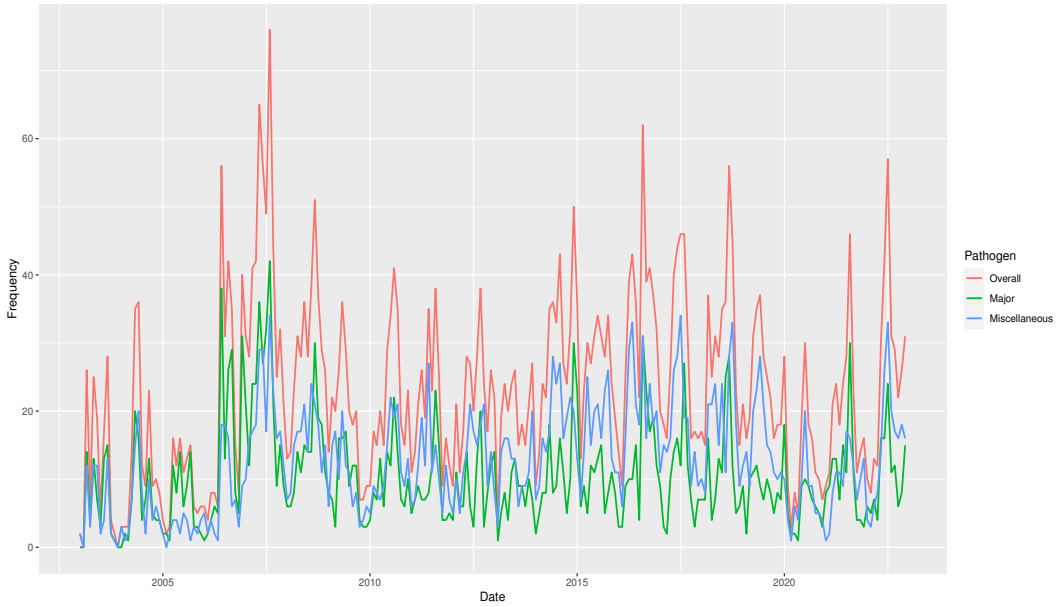

Figure 1: *The trend of monthly food poisoning cases.*

MinT 추정량은 적합값이 0보다 크거나 같아야 한다는 제약조건이 없기 때문에 0 근처의 $\hat{\mu}_{it}$는 음수인 $\tilde{\mu}_{it}$로 유도될 수 있다. 이를 해결하기 위해 $\hat{\mu}_{it}$ 대신 $\log(\hat{\mu}_{it})$를 이용하여 최적결합 추정량을 구하고 이를 지수변환하여 최종결과를 얻는 방법을 생각해 볼 수 있으나 지수변환이 선형이 아니기 때문에 등식조건을 만족하지 않을뿐만 아니라 산술평균이 아닌 기하평균 추정이 되어 이 연구에서 얻고자 하는 결과를 얻지 못한다. 이 논문에서는 음수인 $\tilde{\mu}_{it}$를 0으로 만들고 주요와 기타 평균 발생건수의 계층조정 결과와 일치하도록 다음과 같이 나머지 양수인 자료에 가중치를 곱하여 최종 결과를 유도한다.

$$\tilde{\mu}^{*}_{it} = \tilde{\mu}_{it} I\left(\tilde{\mu}_{it} > 0\right) \frac{\sum_{j=k_1}^{k_2} \tilde{\mu}_{jt}}{\sum_{j=k_1}^{k_2} \tilde{\mu}_{jt} I(\tilde{\mu}_{jt} > 0)} = \tilde{\mu}_{it} I(\tilde{\mu}_{it} > 0) \frac{\tilde{\mu}^{(k)}_{t}}{\sum_{j=k_1}^{k_2} \tilde{\mu}_{jt} I(\tilde{\mu}_{jt} > 0)},$$

여기서 $I(\cdot)$는 지시함수를 의미하며 $i = 1, \ldots, 5$일 때 $k_1 = 1$, $k_2 = 5$, $k = 1$이고 $i = 6, \ldots, 14$일 때 $k_1 = 6$, $k_2 = 14$, $k = 2$이다. 이렇게 재조정하면 모든 원인별 평균발생건수의 추정이나 예측값이 0보다 크거나 같고 계층구조를 만족한다.

3. 실증분석

식중독 발생 자료는 식품의약품안전처 식품안전나라 홈페이지에서 다운로드 받을 수 있으며 인터넷에서 '식중독 통계'를 검색하면 해당 사이트를 찾을 수 있다. 이 사이트에서는 월별/지역별, 월별/원인시설별, 월별/원인물질별로 나누어진 식중독 발생건수와 환자수를 제공한다. 논문을 작성한 시점에서 2022년까지의 자료는 확정된 상태이고 2023년 자료는 잠정, 2024년자료는 신고된 통계였다. 본 연구에서는 확정된 2002년부터 2022년까지의 자료를 분석하였다. Table 1에 표시된 것과 같이 월별/원인물질별 자료에는 서론에서 언급한 5개 주요 원인균과 불명을 포함한 9가지 원인별로 식중독 건수와 환자수를 제공하고 있다.

앞 절에서 설명한 것처럼 원인물질별 식중독 발생건수는 영과잉 포아송 회귀모형과 영과잉 음이항 회귀

Table 2: Goodness of forecast

Pathogen	Poisson						Negative binomial					
	$\sqrt{\text{MSE}}$		MAE		MAX		$\sqrt{\text{MSE}}$		MAE		MAX	
	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$	$\hat{\boldsymbol{\mu}}_t$	$\tilde{\boldsymbol{\mu}}_t^*$
PE	3.20	3.19	2.08	2.07	12.73	12.48	3.25	3.24	2.13	2.14	12.46	12.27
Salmonella	1.89	1.90	1.23	1.23	10.27	10.24	1.93	1.94	1.25	1.25	10.27	10.22
VP	2.43	2.40	1.09	1.06	20.34	20.24	2.43	2.39	1.10	1.09	20.02	19.95
Norovirus	4.34	4.34	2.96	3.03	20.21	20.39	4.41	4.41	2.98	3.05	20.31	20.55
SA	1.64	1.63	1.09	1.10	8.78	8.76	1.67	1.67	1.11	1.12	9.00	9.00
CJ	1.77	1.75	1.10	1.09	9.00	8.86	1.72	1.70	1.06	1.05	9.13	8.97
CP	1.54	1.54	1.04	1.04	6.48	6.48	1.42	1.44	0.96	0.98	6.00	5.97
BC	0.88	0.89	0.64	0.64	4.29	4.29	0.89	0.89	0.65	0.66	4.29	4.29
OB	0.51	0.50	0.24	0.24	3.00	2.99	0.53	0.53	0.25	0.25	3.00	2.98
OV	0.48	0.48	0.32	0.32	2.00	2.00	0.46	0.45	0.31	0.31	1.86	1.86
Protozoa	2.39	2.38	1.98	1.97	4.93	4.88	10.01	3.55	3.71	2.32	73.99	18.05
NT	0.45	0.45	0.31	0.31	2.63	2.64	0.46	0.46	0.30	0.30	2.85	2.85
CS	0.24	0.24	0.08	0.08	1.46	1.46	0.24	0.24	0.07	0.08	1.45	1.45
Unknown	7.11	7.48	5.61	5.91	25.44	24.78	7.34	7.88	5.77	6.15	25.68	24.86
Major	8.04	7.99	5.74	5.62	30.17	31.02	8.18	8.08	5.84	5.71	30.75	31.05
Miscellaneous	9.74	9.14	7.56	7.20	32.74	26.73	10.59	9.77	8.08	7.61	37.70	28.19
Overall	14.84	14.67	11.64	11.46	50.89	52.04	15.52	15.16	12.13	11.83	51.33	52.45

모형으로 분석하고 주요 원인균, 기타 원인, 전체 식중독 발생건수는 포아송 회귀모형과 음이항 회귀모형으로 분석하였다. 설명변수에 평균기온, 평균습도, 평균기온과 평균습도의 상호작용, 일사합, 수온, 전월 발생건수를 포함시켰고 Figure 1에서 보는 것과 같이 계절성을 있는 것으로 나타났으며 전년 동월 발생건수를 추가하였다. 2007년 전과 비교해 2007년 이후에 발생건수가 늘어났으며 Covid 19가 발생하기 시작한 2020년에는 기존에 비해 줄어든 경향이 있으나 본 연구에서는 그에 대한 추가 요인을 반영하지 않고 분석하였다. 전년 동월 발생건수를 설명변수로 사용했기에 분석에서는 설명변수를 완벽히 포함한 2003년 1월 자료부터 2022년 12월 자료를 사용하였다. 원인물질별 발생여부와 발생건수에 유의하게 영향을 주는 설명변수가 다를 수 있기 때문에 AIC 기준에서 변수선택을 하였다. 주요 원인균, 기타 원인, 전체 발생건수에 대해서도 AIC 기준으로 변수선택하여 최종 결과를 유도하였다.

적합력을 확인하기 위해 240개 시점의 자료를 기반으로 $\hat{\boldsymbol{\mu}}_t$와 $\tilde{\boldsymbol{\mu}}_t^*$를 유도하고 $\boldsymbol{y}_t$와 비교하였다. 포아송분포나 음이항분포에 대한 적합력은 아래와 같은 Pearson 잔차나 Anscombe 잔차를 기반으로 비교 분석하는 것이 적절할 수 있으나 영과잉 회귀분석에 따른 계층조정 값 $\tilde{\boldsymbol{\mu}}_t^*$에 0이 존재할 수 있어 분모가 $\sqrt{\hat{y}_{it}}$나 $\hat{y}_{it}^{1/6}$인 잔차로 적합력을 설명하는 것은 적절하지 않다.

$$r_{it}^{(P)} = \frac{y_{it} - \hat{y}_{it}}{\sqrt{\hat{y}_{it}}}, \quad r_{it}^{(A)} = \frac{3(y_{it}^{2/3} - \hat{y}_{it}^{2/3})}{2\hat{y}_{it}^{1/6}}.$$

이 논문에서는 적합력이나 예측력을 비교할 때 일반적으로 사용하는 평균제곱오차(MSE), 평균절대오차(MAE), 최대오차(MAX)로 계층조정을 한 $\tilde{\boldsymbol{\mu}}_t^*$와 하지 않은 $\hat{\boldsymbol{\mu}}_t$를 비교한다.

$$\sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_t (y_t - \hat{y}_t)^2}, \quad \text{MAE} = \frac{1}{n}\sum_t |y_t - \hat{y}_t|, \quad \text{MAX} = \max|y_t - \hat{y}_t|,$$

또한 예측력을 확인하기 위해 2003년 1월 자료부터 2016년 12월까지의 자료를 분석하여 2017년 1월 발생건수에 대한 $\hat{\boldsymbol{\mu}}_t$와 $\tilde{\boldsymbol{\mu}}_t^*$를 유도하고 $\boldsymbol{y}_t$와 비교한다. 이 때 기상자료는 예측 가능하다고 가정하고 2017년 1월의

예측값을 구할 때 해당시점의 기상자료를 설명변수로 사용하였다. 그 다음부터 분석자료를 하나씩 증가시키며 다음 달의 자료를 예측하도록 하였다. 이런 방식으로 얻어진 72개($t = 169, \ldots, 240$) 예측결과를 이용하여 $\sqrt{\text{MSE}}$, MAE, MAX로 예측력을 비교하였다.

분석에서는 R 4.3.2를 이용하였으며 포아송 회귀분석과 음이항 회귀분석은 glm()과 MASS::glm.nb()를, 변수선택은 step()을 사용하였다. 영과잉 포아송회귀분석과 영과잉 음이항회귀분석에는 pscl::zeroinfl()이 사용되었고 mpath::be.zeroinfl()로 변수선택을 하였다. 시계열 계층조정은 hts::hts()와 hts::aggts() 함수를 적용하고 추가적인 부분은 직접 코딩하여 분석하였다. 현재 hts 패키지는 더 이상 업데이트 하지 않고 관리차원으로 CRAN에서 제공하는데 관련 연구를 위한 fable 패키지를 제공하고 있으니 참고하기 바란다.

Table 1은 240개 시점의 적합값과 관측값을 비교한 결과로 음이항 분포를 기반으로 분석한 결과보다는 포아송분포를 기반으로 분석한 결과가 전반적으로 조금 더 좋은 것으로 나타났으며 원인물질별 분석에서는 계층조정을 한 것과 하지 않은 것 간 차이는 거의 없었으며 주요(Major), 기타(Miscellaneous), 전체(Overall)에서는 계층조정의 결과가 더 우수한 것으로 나타났다. 또한 원일물질별 발생건수 적합값이 주요나 기타 발생건수의 적합값보다 큰 경우, 즉 ($i = 1, \ldots, 5, j = 1$)과 ($i = 6, \ldots, 14, j = 2$)에 대해 $\hat{\mu}_{it} > \hat{\mu}_t^{(j)}$인 경우가 포아송은 23건, 음이항은 27건인 것으로 나타났다.

Table 2는 2017년 1월부터 2022년 12월까지 72개 시점에 대한 예측값과 실제 관측값을 비교한 결과로 Table 1의 결과에 비해 상대적으로 예측력이 떨어지는 것을 볼 수 있다. 적합값의 비교에서와 유사하게 포아송분포를 기반으로 한 예측이 음이항 분포 기반한 예측보다 조금 좋은 것으로 나타났고 주요, 기타, 전체에 대한 예측력에서 계층조정한 결과가 하지 않은 결과보다 대체로 우수한 것으로 볼 수 있다. 포아송이나 음이항 회귀분석을 기반으로 예측과정에서 추정된 선형예측값(linear predictor)을 지수변환하여 평균을 추정하게 되는데 이 과정에서 비정상적으로 큰 예측값이 출력되는 경우가 가끔 있다. 예를 들어, 음이항 분포를 기반으로 분석한 원충(Protozoa)의 비계층조정 MAX 값이 73.99로 나왔는데 이는 2019년 8월의 예측값이고 해당 시점의 관측값은 0건으로 상당히 과대 추정한 것으로 나타났다. 또한 그 시점에서 기타 원인을 모두 합한 자료로 예측한 기타(Miscellaneous)의 예측값이 18.76으로 계층구조에서는 설명하기 어려운 결과가 나왔다. 이에 반해 계층조정한 값은 18.05로, 기타 원인에 의한 예측값은 24.9으로 보정 되었다. 적합 때와 마찬가지로 $\hat{\mu}_{i,t} > \hat{\mu}_t^{(j)}$인 경우가 포아송은 5건, 음이항은 7건 발생하였다.

4. 결론

이 논문에서는 원인물질별 식중독 발생건수를 영과잉 포아송 회귀모형과 영과잉 음이항 회귀모형으로 분석하고 주요, 기타 및 전체 발생건수를 포아송 회귀모형과 음이항 회귀모형으로 분석하였다. 여기서 강조하고 싶은 것은 원인물질별, 주요, 기타 및 전체 식중독 발생건수를 어떤 모형으로 분석했는지가 아니라 계층적 구조를 가지는 이들 자료의 적합값과 예측값들에 대해 어떻게 하면 계층구조를 만족하게 할 것인지에 대한 방법론이다. 여기에서는 최적결합, 특히 Wickramasuriya 등 (2019)가 제안한 MinT 추정량을 이용했는데 MinT 추정의 경우 조정된 값이 음수가 될 수 있다. 이 경우 음수를 0으로 처리하고 다른 최하위 변수의 적합값 또는 예측값에 가중치를 곱해 최종적으로 계층구조를 만족시키는 방법에 대해 알아보았다. 실제로 발생 빈도와 같이 0 또는 양의 관측값만 존재하면서 계층구조를 가지는 자료가 상당히 많은데 이러한 경계 조건을 가지는 경우에도 적용할 수 있는 계층 시계열 예측 연구가 필요하며 활용도 높은 연구가 될 것이다.

이 논문에서는 월별/원인물질별 식중독 자료를 분석하였으며 설명변수에는 전국 평균 기상 자료가 사용되었다. 우리나라는 크지 않지만 바다, 평야와 산맥 등이 혼재되어 지역에 따라 기후 차이가 있으며 원인물질의 활성도에도 차이가 있기 때문에 보다 직접적으로 기상자료가 식중독에 어떻게 영향을 주는지 파악하기 위해서는 지역별 분석이 필요하다. 식품안전나라 홈페이지에서 제공하는 월별/지역별 식중독 자료를 동일한 방법으로 분석하여 예측값을 유도할 수 있으나 국가 차원에서의 식중독 예방 대책은 원인물질별로 세우는 것

이 일반적이기 때문에 원인물질별 식중독 분석을 기반으로 지역적 요소를 반영하는 것이 적절하다. 또한 월간 기상 예보는 정확성이 낮지만 주간 기상 예보는 정확도가 상대적으로 높기 때문에 주별/원인물질별/지역별 식중독 자료를 수집할 수 있으면 좀 더 실용적인 식중도 예측을 할 수 있을 것으로 생각된다.

References

Agarwal DK, Gelfand AE, and Citron-Pousty S (2002). Zero-inflated models with application to spatial count data, *Environmental and Ecological Statistics*, **9**, 341–355.

Choi K, Kim B, Bae W, Jung W, and Cho Y (2008). Developing the index of foodborne disease occurrence, *The Korean Journal of Applied Statistics*, **21**, 649–658.

Dunn DM, Williams WH, and DeChaine TL (1976). Aggregate versus subaggregate models in local area forecasting, *Journal of the American Statistical Association*, **71**, 68–71.

Fleury M, Charron DF, Holt JD, Allen OB, and Maarouf AR (2006). A time series analysis of the relationship of ambient temperature and common bacterial enteric infections in two Canadian provinces, *International Journal of Biometeorology*, **50**, 385–391.

Hyndman RJ, Ahmed RA, Athanasopoulos G, and Shang HL (2011). Optimal combination forecasts for hierarchical time series, *Computational Statistics and Data Analysis*, **55**, 2579–2589.

Jung HS., Kim BJ, Cho S, and Yeo, IK (2012). Analysis of food poisoning via zero inflation models, *The Korean Journal of Applied Statistics*, **25**, 859–864.

Jeong MS and Oh SS (2009). *Study on the Impact Analysis and Control System of Foodborne Disease Outbreak due to Climate Change*, Korea Food & Drug Administration, Cheongju-si.

Kim S, Seong B, Choi YG, and Yeo IK (2022). A study on time series linkage in the household income and expenditure survey, *The Korean Journal of Applied Statistics*, **35**, 553–568.

Lambert D (1992). Zero-inflated Poisson regression models with an application to defects in manufacturing, *Technometrics*, **34**, 1–14.

Orcutt GH, Watts HW, and Edwards JB (1968). Data aggregation and information loss, *The American Economic Review*, **58**, 773–787.

Shlifer E and Wolff RW (1979). Aggregation and proration in forecasting, *Management Science*, **25**, 594–603.

Wickramasuriya SL, Athanasopoulos G, and Hyndman RJ (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization, *Journal of the American Statistical Association*, **114**, 804–819

Yeo IK (2012). Models for forecasting food poisoning occurrences, *Journal of the Korean Data & Information Science Society*, **23**, 1117–1125.

Yeo IK (2013). Prediction of the number of food poisoning occurrences by microbes, *The Korean Journal of Applied Statistics*, **26**, 923–932.

Zhang Y, Bi P, and Hiller JE (2008). Climate variations and salmonellosis transmission in Adelaide, South Australia: A comparison between regression models, *International Journal of Biometeorology*, **52**, 179–187.

Zhang Y, Bi P, Hiller JE, Sun Y, and Ryan P (2007). Climate variations and bacillary dysentery in northern and southern cities of China, *The Journal of Infection*, **55**, 194–200.

Received March 07, 2024; Revised April 03, 2024; Accepted April 06, 2024

식중독 발생 건수에 대한 계층 시계열 예측

여인권[1,a]

[a]숙명여자대학교 통계학과

요 약

이 연구에서는 식중독 발생건수를 원인물질별로 나눈 자료와 합한 자료를 별개로 분석하여 예측값을 유도한 후 계층구조를 만족하도록 하는 계층 시계열 예측에 대해 알아본다. 원인물질별 식중독 방생건수는 영과잉 포아송 회귀모형과 음이항 회귀모형으로 분석하고 합한 식중독 발생건수 포아송 회귀모형과 음이항 회귀모형으로 분석한다. 계층 시계열 예측을 위해 최적결합 중 하나인 Wickramasuriya 등 (2019)의 MinT 추정이 사용되었다. 계층조정 과정에서 발생한 음의 예측값은 0으로 수정하고 나머지 최하위 변수에 가중치를 곱해 계층구조를 만족시킨다. 실증분석 결과를 보면 원인물질별 예측에서는 계층조정을 한 결과와 하지 않은 결과에 차이가 거의 없었으나 주요, 기타 및 전체에 대한 예측에서는 계층조정 한 결과가 대체로 우수한 것으로 나타났다. 중요한 것은 계층조정을 하지 않으면 최하위 변수의 예측빈도가 주요나 기타의 예측빈도 보다 큰 경우도 발생하지만 제안된 방법을 적용하면 계층구조를 이루는 예측값을 얻을 수 있다.

주요용어: 계층 시계열 예측, 영과잉 회귀모형, 음이항 회귀모형, 최적 결합, 포아송 회귀모형

[1](04310) 서울특별시 용산구 청파로47길 100, 숙명여자대학교 통계학과. E-mail: inkwon@sookmyung.ac.kr