The Korean Journal of Applied Statistics (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.2.147
33(2), 147–160

Automatic order selection procedure for count time series models

Yunmi Ji[a] · Byeongchan Seong[a,1]

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

(Received December 30, 2019; Revised January 12, 2020; Accepted January 12, 2020)

Abstract

In this paper, we study an algorithm that automatically determines the orders of past observations and conditional mean values that play an important role in count time series models. Based on the orders of the ARIMA model, the algorithm constitutes the order candidates group for time series generalized linear models and selects the final model based on information criterion among the combinations of the order candidates group. To evaluate the proposed algorithm, we perform small simulations and empirical analysis according to underlying models and time series as well as compare forecasting performances with the ARIMA model. The results of the comparison confirm that the time series generalized linear model offers better performance than the ARIMA model for the count time series analysis. In addition, the empirical analysis shows better performance in mid and long term forecasting than the ARIMA model.

Keywords: count time series, automatic algorithm, time series generalized linear model, ARIMA model

1. 서론

시계열 자료는 종종 횟수(count data)로 구성된다. 예를 들면, 월별 승합차 운전자의 사상자 수, 주별 자동차 서비스 부품의 재고량, 일별 말라리아 감염 환자 수 등 다양한 분야에서 횟수로 구성된 시계열 자료를 접할 수 있다. 이와 같이 일정한 시간 간격 동안 발생하는 사건의 수에 관련된 시계열 자료를 계수형 시계열 자료(count time series)라고 한다. 계수형 시계열 자료의 가장 큰 특징은 음이 아닌 정수의 값(non-negative integer value)을 갖는다는 것이다. 또한, 흔히 발생하지 않는 사건의 경우 영이 많이 관측되는 영과잉 시계열의 형태를 띄기도 한다.

대표적인 시계열 분석 모형인 자기회귀누적이동평균(autoregressive integrated moving average; ARIMA) 모형을 계수형 시계열 자료 분석에 사용할 수 있지만 한계점을 갖는다. 첫째, ARIMA 모형은 오차항의 분포가 정규분포를 따른다고 가정하며 이 분포의 형태는 대칭적인 종 모양이지만 계수형 시계열 자료의 경우, 특히 값이 적은 경우, 경험적 분포가 다소 편향된 형태를 갖기 때문에 적절하지 않다. 일반적으로 편향된 자료에 대해 로그 변환을 한 후 정규분포 가정을 하는 경우도 있지만 0을 많

This research was supported by the Chung-Ang University Research Scholarship Grants in 2018 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

이 포함하고 있는 계수형 시계열 자료인 경우에는 적용이 어렵다. 둘째, ARIMA 모형의 표본 공간은 $(-\infty, \infty)$의 범위를 갖는 실수(real-valued)이다. 즉, 계수형 시계열 자료에 ARIMA 모형을 적용할 경우 이산형(discrete) 표본 공간의 특성이 고려되지 않는다. 따라서 ARIMA 모형은 계수형 시계열 자료 분석에 적절하지 않으며, 최근 ARIMA 모형이 갖는 문제점을 개선하고 계수형 시계열 자료의 특징을 고려한 다양한 모형이 활발하게 연구되고 있다.

계수형 시계열의 모형은 관측치가 음수가 아닌 정수임을 고려해야 하며, 관측치 사이의 의존성을 적절히 포착해야 한다. 가장 편리하고 유연한 접근 방법은, Fahrmeir과 Tutz (2001, 6장) 그리고 Kedem과 Fokianos (2002, 1-4장)가 제안한 과거 정보에 대한 관측치를 조건부로 모형화하기 위해 일반화 선형 모형(generalized linear model; GLM)을 사용하는 것이다. 이 모형은 계수형 자료에 대한 적절한 분포와 연결 함수를 선택하여 사용한다. 또다른 중요한 방법은 Weiß(2008, 2-3장)이 소개한 정수형 자기회귀이동평균(integer autoregressive moving average; INARMA) 모형이다. 이는 'thinning operator'에 기초하는데, 일반적인 자기회귀이동평균 모형의 구조를 모방한다.

시계열 일반화 선형 모형은 모형에 사용되는 과거 관측값의 차수와 과거 조건부 평균값의 차수를 결정해야한다. 단순히 시계열 그림을 보고 적절한 차수를 결정하는 것은 어렵기 때문에 대안으로 자기상관함수(autocorrelation function; ACF)를 고려하여 높은 자기 상관성을 갖는 시차를 모형의 차수로 정할 수 있다. 그러나, 현재까지 모형의 차수를 결정하는 자동화 알고리즘이 존재하지 않기 때문에 불편한 식별(identification) 작업을 통해야만 한다. 유사한 방법을 통해 모형의 차수를 결정하는 ARIMA 모형의 경우 자동화 차수 결정 알고리즘이 이미 존재하지만 차분의 개념과 계절형 승법 구조를 시계열 일반화 선형 모형에 그대로 반영하기 어려워 ARIMA 모형의 차수 결정을 바로 이용할 수 없다. 따라서 본 논문에서는 이를 고려하여 시계열 일반화 선형 모형의 차수 후보군을 만들고, 후보군의 조합을 이용하여 적합한 모형 중 AIC가 가장 작은 모형을 최종 모형으로 선택하는 ARIMA 모형 차수 결정에 기반한 자동 차수 결정 알고리즘을 고안하였다.

본 논문에서는 시계열 일반화 선형 모형의 차수 결정 알고리즘 고안과 시뮬레이션 및 실증분석을 통하여 ARIMA 모형과의 예측 성능을 비교하였다. 본 논문은 총 5장으로 구성되어 있으며 2장에서는 시계열 일반화 선형 모형 및 추정 방법을 설명하고, 3장에서는 차수를 결정해주는 자동화 알고리즘 소개와 ARIMA 모형과의 예측 성능 비교를 위한 시뮬레이션을 진행한다. 4장에는 국내 살인사건 발생 건수 자료를 이용하여 모형의 적합 및 예측 성능을 비교 분석하며 마지막 5장에서는 결론을 맺는다.

2. 시계열 일반화 선형 모형

2.1. 기본 모형

시계열 일반화 선형 모형은 조건부 평균이 시간-변동 공변량, 과거의 관측값 그리고 과거의 조건부 평균값에 따라 변하는 구조를 가지는 모형이다. 원 시계열 자료를 $\{Y_t : t \in \mathbb{N}\}$, 시간-변동 공변량 벡터를 $X_t = (X_{t,1}, \ldots, X_{t,r})^T$, 과거의 정보 F_{t-1}에 대하여 Y_t의 조건부 평균을 $E(Y_t \mid F_{t-1}) = \lambda_t$라고 할 때, 시계열 일반화 선형 모형은 다음과 같다.

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-k}) + \sum_{l=1}^{q} \alpha_l g(\lambda_{t-l}) + \eta^T X_t, \tag{2.1}$$

여기서 g와 $\tilde{g}$는 각각 연결 함수와 변환 함수이며, 과거의 정보 F_t는 $t+1$시점의 공변량 정보를 포함하여 $\{Y_t, \lambda_t, X_{t+1} : t \in \mathbb{N}\}$의 결합 확률 과정이다.

임의의 과거 관측을 선형 모형에 반영하기 위해 집합 $P = \{1, \ldots, p\}$를 정의해야 한다. 이때, 적절한

p를 선택하고, 조건부 평균에 영향을 주지 않는 일부 과거 시차에 대한 관측 Y_{t-k}의 모수를 0으로 설정해야 한다. 마찬가지로 과거 조건부 평균을 선형 모형에 반영하기 위해 집합 $Q = \{1, \ldots, q\}$를 정의하고, 적절한 q의 선택과 일부 과거 시차에 대한 조건부 평균 λ_{t-l}의 모수를 0으로 설정하는 것이 필요하다. 구체적인 모형의 차수(즉, P와 Q의 집합)의 결정을 위하여 원 시계열 자료의 자기상관함수를 고려할 수 있다.

2.1.1. 분포 가정 시계열 일반화 선형 모형에서 과거의 정보 F_{t-1}에 대하여 Y_t의 분포는 포아송 분포 또는 음이항 분포를 가정한다. 먼저, 모형 (2.1)과 함께 포아송 분포를 가정한 경우 다음과 같이 표현될 수 있다.

$$\begin{aligned} Y_t \mid F_{t-1} &\sim \text{Poisson}\left(\lambda_t\right), \\ P\left(Y_t = y \mid F_{t-1}\right) &= \frac{\lambda_t \exp\left(-\lambda_t^y\right)}{y!}, \quad y = 0, 1, \ldots. \end{aligned} \tag{2.2}$$

이 경우 $\text{Var}(Y_t \mid F_{t-1}) = E(Y_t \mid F_{t-1}) = \lambda_t$이다.

음이항 분포는 조건부 분산이 조건부 평균보다 클 수 있도록 과산포(overdispersion)의 개념을 도입한다. 음이항 분포를 가정한 경우 평균의 관점에서 과산포 모수인 $\phi \in (0, \infty)$가 추가적으로 모수화되며 다음과 같이 표현된다.

$$\begin{aligned} Y_t \mid F_{t-1} &\sim \text{NegBin}\left(\lambda_t, \phi\right), \\ P\left(Y_t = y \mid F_{t-1}\right) &= \frac{\Gamma(\phi + y)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\phi}{y+1}\right)^{\phi} \left(\frac{\lambda_t}{\phi + \lambda_t}\right), \quad y = 0, 1, \ldots. \end{aligned} \tag{2.3}$$

이 경우 $\text{Var}\left(Y_t \mid F_{t-1}\right) = \lambda_t + \lambda_t/\phi$이며 조건부 분산은 λ_t에 따라 2차적으로 증가한다. 포아송 분포는, $\phi \to \infty$일 때 음이항 분포의 극한 분포이다.

2.1.2. 연결 함수 일반적으로 시계열 일반화 선형 모형의 연결 함수로 항등 함수와 로그 함수를 고려한다. 먼저, 연결 함수 g와 변환 함수 $\tilde{g}$가 모두 항등함수인 경우, 즉 $g(x) = \tilde{g}(x) = x$, 모형 (2.1)은 다음과 같다.

$$\lambda_t = \beta_0 + \sum_{k=1}^{p} \beta_k Y_{t-k} + \sum_{l=1}^{q} \alpha_l \lambda_{t-l} + \eta^T X_t. \tag{2.4}$$

추가적으로 $\eta = 0$과 $Y_t \mid F_{t-1}$의 분포를 포아송 분포로 가정할 때 모형 (2.4)를 차수가 p, q인 integer-valued GARCH 모형, INGARCH(p, q)으로 부른다. 이 모형은 자기상관 조건부 포아송(autoregressive conditional Poisson; ACP) 모형으로도 알려져 있으며 자세한 내용은 Ferland 등 (2006)과 Fokianos 등 (2009)를 참고하라.

연결 함수가 로그 함수인 경우 연결 함수는 $g(x) = \log(x)$, 변환 함수는 $\tilde{g}(x) = \log(x+1)$이며 다음과 같다.

$$\log(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \log\left(Y_{t-k} + 1\right) + \sum_{l=1}^{q} \alpha_l \log(\lambda_{t-l}) + \eta^T X_t. \tag{2.5}$$

자세한 내용은 Fokianos와 Tiøstheim (2011)을 참고하라.

2.2. 모수 추정 방법

모형 (2.1)을 적합시키기 위하여 준조건부 최대 가능도(quasi-conditional maximum likelihood) 추정을 사용한다. 모수 벡터를 $\theta = (\beta_0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta_1, \ldots, \eta_r)^T$라고 할 때, 연결 함수가 항등 함수인 경우의 모수 공간은 다음과 같다.

$$\Theta = \left\{ \theta \in \mathbb{R}^{p+q+r+1} : \beta_0 > 0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \eta_1, \ldots, \eta_r \geq 0, \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l < 1 \right\}. \quad (2.6)$$

조건부 평균 λ_t가 양수가 되기 위해 상수 β_0은 양수이고, 다른 모든 모수는 음수가 아닌 값을 가져야 하며, 정상성(stationary)과 에르고딕성(ergodic)을 위해 마지막 조건을 만족해야 한다. 자세한 내용은 Tiøstheim (2015)과 Doukhan 등 (2012)을 참고하라. 또한, 연결 함수가 로그 함수인 경우 일반화 선형 모형의 모수 공간은 다음과 같다.

$$\Theta = \left\{ \theta \in \mathbb{R}^{p+q+r+1} : |\beta_1|, \ldots, |\beta_p|, |\alpha_1|, \ldots, |\alpha_q|, \left| \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l \right| < 1 \right\}. \quad (2.7)$$

과거의 정보 F_{t-1}에 대하여 Y_t의 분포를 음이항 분포로 가정한 경우 회귀 모수 θ의 추정은 과산포 모수 ϕ에 의존하지 않는다. 따라서, 회귀 모수 θ를 추정하기 위해 포아송 가능도 기반의 준최대 가능도(quasi-maximum likelihood) 방법의 사용을 가능하게 하며, 과산포 모수 ϕ는 회귀 모수 θ를 추정한 후 별개로 추정한다. 자세한 내용은 Christou와 Fokianos (2014)를 참고하라. 관측 벡터 $y = (y_1, \ldots, y_n)^T$에 대하여 조건부 준로그 가능도 함수(conditional quasi log-likelihood function)는 다음과 같다.

$$l(\theta) = \sum_{t=1}^{n} \log p_t(y_t; \theta) = \sum_{t=1}^{n} (y_t \ln(\lambda_t(\theta)) - \lambda_t(\theta)) \quad (2.8)$$

여기서 $p_t(y_t; \theta) = P(Y_t = y \mid F_{t-1})$는 식 (2.2)에서 정의한 포아송 분포의 확률 질량 함수이다. 조건부 평균 λ_t는 모든 t에 대하여 θ에 관한 함수이므로 $\lambda_t(\theta)$로 표현된다.

θ의 준최대 가능도 추정량(quasi maximum likelihood estimation; QMLE) $\hat{\theta}_n$은 비선형 제약 최적 문제(non-linear constrained optimization problem)의 해이며 다음과 같이 표현할 수 있으며 자세한 내용은 Liboschik 등 (2017)을 참고하라.

$$\hat{\theta} := \hat{\theta}_n = \arg\max_{\theta \in \Theta} l(\theta). \quad (2.9)$$

2.3. 예측 방법

평균 제곱 오차의 측면에서 n시점까지의 과거의 정보 F_n이 주어졌을 때 최적의 1시점 이후 미래 예측치 $\hat{Y}_{n+1}$은 조건부 평균 λ_{n+1}이다. 모형 구성에 따라 $\hat{Y}_{n+1}$의 분포는 각각 평균 λ_{n+1}을 갖는 포아송 분포 또는 음이항 분포이다. Y_{n+h}에 대한 미래 예측치 $\hat{Y}_{n+h}$는 반복적인 1시점 이후의 예측치에 의해 얻어지며 비관측값 $Y_{n+1}, \ldots, Y_{n+h-1}$은 각각의 1시점 이후의 예측치로 대체되어 사용된다. h시점 이후의 예측치 $\hat{Y}_{n+h}$의 분포는 통계적으로 알려지지 않았지만 수리적으로 모수적 부스트랩(parametric bootstrap)을 통해 근사시킬 수 있다. 자세한 내용은 Liboschik 등 (2017)을 참고하라.

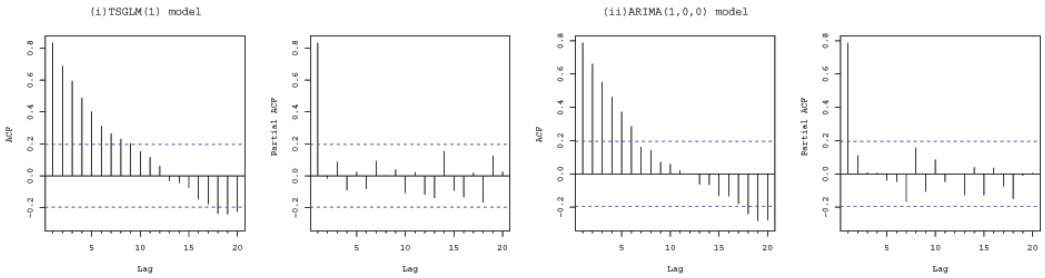

Figure 3.1. ACF and PACF plots of TSGLM(1)(0) and ARIMA$(1, 0, 0)$.

3. 차수결정 자동화 알고리즘

3.1. 알고리즘

시계열 일반화 선형 모형은 과거 시차의 관측값과 조건부 평균을 모형에 반영하기 위하여 적절한 차수가 필요하며, 이를 위하여 대안으로 관측된 자료의 자기상관함수를 이용할 수 있다. 유사한 방법을 통해 모형의 차수를 결정하는 ARIMA 모형의 경우, 자동화 차수 결정 알고리즘이 R패키지 forecast (Hyndman과 Khandakar, 2008)의 auto.arima 함수에 구현되어 있다. 그러나, 시계열 일반화 선형 모형은 ARIMA 모형과 달리 차분의 개념이 없고 ARIMA 모형의 계절형 승법 구조를 가지고 있지 않으므로 해당 알고리즘을 그대로 이용하는 것은 불가능하다.

본 논문에서는 R 패키지 tscount (Liboschik 등, 2017)의 tsglm 함수에 의하여 계산되는 정보량 기준값(Akaike information criterion; AIC)을 이용하여 차수 결정 자동화 알고리즘을 고안한다. 참고로 tsglm 함수에서 차수 P와 Q에 해당하는 인자는 각각 past_obs과 past_mean이다. $Y_t \mid F_{t-1}$의 분포로서 포아송 분포, 연결 함수로 항등함수를 사용할 때, 이러한 모형을 TSGLM$(P)(Q)$로 표시하기로 한다. 부분 모형을 위하여 차수를 부분적으로 적용할 때는, TSGLM$(P_1, P_2, \ldots)(Q_1, Q_2, \ldots)$로 표시한다.

먼저, 시계열 일반화 선형 모형의 차수에 따른 자기상관함수와 편자기상관함수(partial ACF; PACF)의 특징을 살펴보기 위해 대표적인 차수에 의해 생성된 계수형 자료의 자기상관함수와 편자기상관함수 그림을 살펴보고자 한다. 그리고, 이것을 ARIMA 모형의 차수에 따른 형태와 비교한다.

Figure 3.1은 TSGLM(1)(0)과 ARIMA$(1, 0, 0)$에서 각각 100개의 관측치로 구성된 시계열 자료를 생성하여 자기상관함수와 편자기상관함수 그림을 나타낸 것이다. 두 모형 모두에서 자기상관함수는 지수적으로 감소하고 있으며, 편자기상관 함수는 lag 1에서 큰 자기 상관성을 갖는 것을 확인할 수 있다.

Figure 3.2는 시계열에 계절성이 반영된 경우로서 TSGLM$(1, 12)(0)$와 ARIMA$(1, 0, 0)(1, 0, 0)_{s=12}$의 자기상관함수와 편자기상관함수 그림을 나타낸 것이다. 두 모형 모두에서 자기상관함수 및 편자기상관함수가 12 주기의 시차들에서 큰 자기 상관성을 갖는 것을 확인할 수 있다.

마지막으로 Figure 3.3은 TSGLM(1)(1)와 ARIMA$(1, 0, 1)$의 자기상관함수와 편자기상관함수 그림을 나타낸 것이다. 두 모형 모두에서 자기상관함수 및 편자기상관함수가 1 시차에서 큰 자기 상관성을 갖는 것을 확인할 수 있다. 따라서 past_obs 차수만 있는 경우, past_obs에 계절 차수를 포함하는 경우, past_obs와 past_mean의 차수를 모두 포함 하는 경우의 세 가지 유형의 시계열 일반화 선형 모형을 통해 ARIMA 모형과 유사한 형태의 자기상관함수와 편자기상관함수를 갖는 것으로 나타났으며, 특히 ARIMA 모형의 AR 차수는 past_obs에 대응되고 MA 차수는 past_mean에 대응되는 것을 확인할 수 있다. 하지만 TSGLM에는 ARIMA 모형의 차분 횟수에 대응되는 인자가 없기 때문에 ARIMA 차수

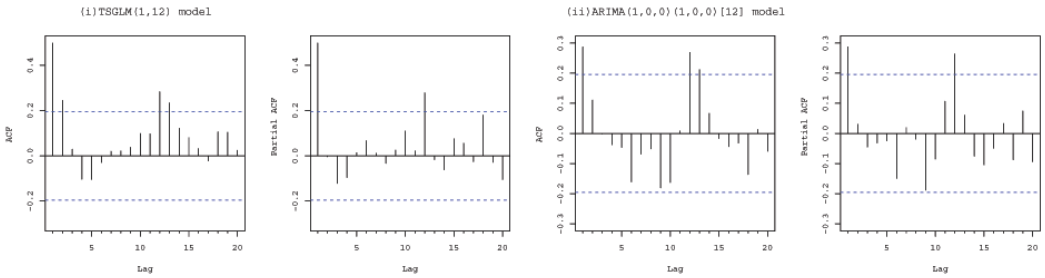

Figure 3.2. ACF and PACF plots of TSGLM$(1, 12)(0)$ and ARIMA$(1, 0, 0)(1, 0, 0)_{s=12}$.

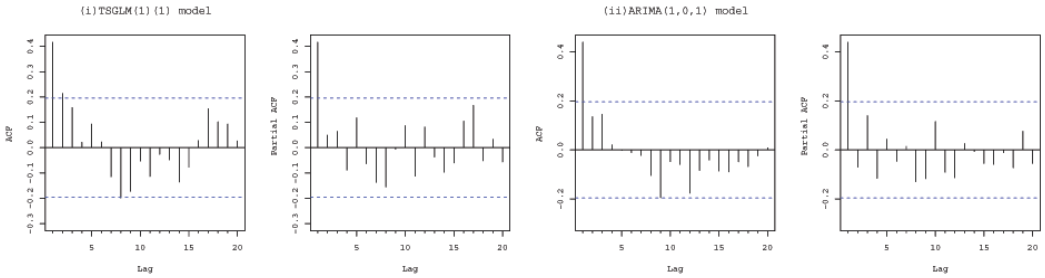

Figure 3.3. ACF and PACF plots of TSGLM$(1)(1)$ and ARIMA$(1, 0, 1)$.

결정을 그대로 따를 수 없다.

본 논문에서 제안하는 자동화 알고리즘은 auto.arima에 의해 결정된 ARIMA 모형의 차수를 이용하여 TSGLM 모형의 차수 후보군을 만드는 것이다. ARIMA 모형의 비계절항에서 AR 차수에 차분 횟수를 포함시키는 경우 차분 횟수만큼 과거 관측의 시차가 증가한다. 따라서 AR 차수와 차분 횟수를 더한 값을 past_obs의 후보 차수로 고려하고, MA 차수를 past_mean의 후보 차수로 고려한다. 강한 자기 상관성을 갖는 차수의 경우 주변 차수에서도 자기 상관성이 높게 나타나기 때문에 각 차수에서 ± 1값도 추가로 고려한다. TSGLM 모형에 계절 효과를 추가하기 위하여, ARIMA 모형의 계절항도 비계절항과 동일한 과정으로, TSGLM 모형의 차수 후보군을 만든다. 이때 계절 주기 차수 직전의 시차에서도 자기 상관성이 높게 나타나기 때문에 후보 차수로 고려한다. 다음으로 past_obs과 past_mean의 후보 차수군을 조합하여 TSGLM 모형을 적합하고, 그 중 AIC가 가장 작은 모형을 최종 모형으로 선택한다. 이상을 요약하면 다음과 같다.

1) auto.arima에 의해 결정된 ARIMA 모형의 차수를 이용하여 TSGLM 모형의 기준 차수를 정한다.
2) AR 차수와 차분 횟수를 합한 값을 past_obs, MA 차수를 past_mean 차수에 각각 대응시킨다.
3) 2) 단계에서 정한 차수에 ± 1값도 TSGLM의 후보 차수로 추가한다.
4) 1)–3) 단계에서 정해진 past_obs와 past_mean의 후보 차수군의 조합을 이용하여 TSGLM 모형을 적합한다.
5) 후보 차수의 조합 중에서 AIC가 가장 작은 모형을 최종 모형으로 선택한다.

3.2. 예측 성능

본 절에서는 3.1절에서 고안한 차수 결정 자동화 알고리즘의 예측 성능을 검토한다. 예측 성능 비교를

Table 3.1. Simulation results in case of the Poisson distribution with the identity link

True model	$N = 50$		$N = 100$		$N = 200$		$N = 400$	
	ARIMA	TSGLM	ARIMA	TSGLM	ARIMA	TSGLM	ARIMA	TSGLM
				mean_RMSE				
TSGLM(1)(0)	3.6291	**3.3024**	3.7493	**3.4573**	3.9086	**3.5676**	3.7709	**3.6172**
TSGLM(2)(0)	3.0712	**2.9188**	3.4334	**3.2656**	3.7314	**3.4937**	3.7467	**3.5664**
TSGLM(3)(0)	3.0715	**2.9633**	3.2861	**3.1533**	3.5232	**3.3651**	3.7380	**3.5541**
TSGLM(1)(1)	2.7361	**2.5855**	2.6946	**2.6169**	2.7744	**2.6494**	2.7373	**2.6748**
TSGLM(2)(1)	2.5554	**2.4973**	2.6352	**2.5299**	2.7118	**2.5726**	2.7040	**2.5956**
TSGLM(1)(2)	2.5873	**2.4619**	2.6340	**2.5431**	2.6402	**2.5411**	2.6315	**2.5650**
				mean_MAE				
TSGLM(1)	3.1352	**2.8330**	3.1296	**2.8667**	3.2093	**2.8991**	3.0189	**2.8817**
TSGLM(2)	2.5565	**2.4143**	2.8098	**2.6611**	3.0195	**2.8147**	2.9868	**2.8371**
TSGLM(3)	2.5337	**2.4332**	2.6483	**2.5278**	2.8140	**2.6856**	2.9583	**2.8090**
TSGLM(1)(1)	2.2978	**2.1524**	2.2194	**2.1461**	2.2471	**2.1350**	2.1950	**2.1419**
TSGLM(2)(1)	2.1246	**2.0728**	2.1614	**2.0675**	2.2050	**2.0781**	2.1712	**2.0757**
TSGLM(1)(2)	2.1661	**2.0489**	2.1693	**2.0825**	2.1331	**2.0447**	2.1109	**2.0545**

위하여 대조 모형으로 ARIMA 모형을 사용하였다. 시뮬레이션을 위해 포아송 분포를 가정하는 시계열 자료의 생성은 R의 tscount 패키지의 tsglm.sim 함수를 사용하였고, 정규분포를 가정하는 시계열 자료의 생성은 R의 forecast 패키지의 arima.sim 함수를 사용하였다. 또한, 시계열 일반화 선형 모형 적합과 ARIMA 모형 적합은 각각 R의 tscount 패키지의 tsglm 함수와 R의 forecast 패키지의 auto.arima 함수를 사용하였다. 모형의 예측 성능을 비교하는 지표는 root mean squared error (RMSE), mean absolute error (MAE)를 사용하였다. 실제 시계열 자료 y_i와 예측값 $\hat{y}_i$에 대하여 미래 시점 h까지의 RMSE와 MAE는 다음과 같이 계산한다.

$$\text{RMSE} = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(y_i - \hat{y}_i)^2}, \tag{3.1}$$

$$\text{MAE} = \frac{1}{h}\sum_{i=1}^{h}|y_i - \hat{y}_i|\,. \tag{3.2}$$

3.2.1. 내재적 모형이 일반화 선형 모형일 경우 내재적 모형(underlying or true model)을 시계열 일반화 선형 모형으로 가정하여, past_obs 차수만 있거나 past_obs과 past_mean 차수가 모두 있는 다음의 6가지 모형들을 고려하였다.

- TSGLM(1)(0), TSGLM(2)(0), TSGLM(3)(0),
- TSGLM(1)(1), TSGLM(1)(2), TSGLM(2)(1).

각 모형에서 N = 50, 100, 200, 400개의 관측치를 생성하였으며, 관측치의 80%는 훈련자료로 사용하여 시계열 일반화 모형과 ARIMA 모형을 적합하였고, 관측치의 20%는 검증자료로 사용하여 두 모형의 RMSE와 MAE를 계산하였다. 동일한 실험을 1,000번 반복함으로써 얻어진 RMSE_i와 MAE_i ($i = 1, \ldots, 1000$)의 평균값은 Table 3.1를 통해 확인할 수 있다. 모든 상황에서 ARIMA 모형에 비해 일반화 선형 모형의 평균 RMSE와 평균 MAE의 값이 더 작게 나타났다. 따라서 내재적 모형이 시계열 일반화 선형 모형이고 차수가 ARIMA 모형의 비계절형 경우처럼 단순한 경우 생성된 시계열 자료의 길

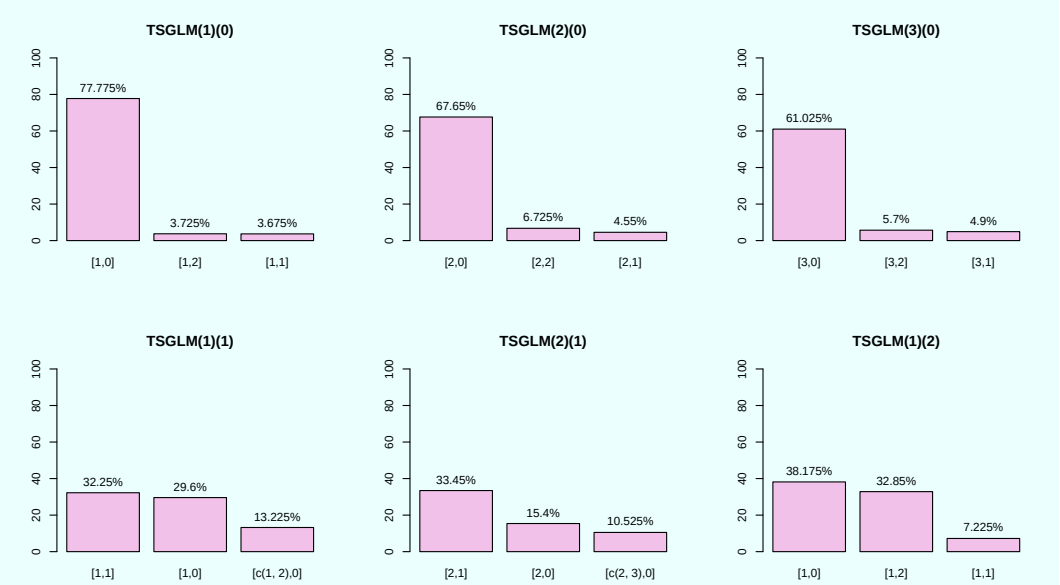

Figure 3.4. Percentage barplot for comparing estimated orders with true values.

이와 상관없이 ARIMA 모형의 예측 성능보다 시계열 일반화 선형 모형의 성능이 더 우수한 것을 확인할 수 있다.

또한 3.1절의 자동화 알고리즘의 차수 결정 성능을 살펴보기 위하여, 일반화 선형모형의 추정된 차수와 참값을 비교하였다. Figure 3.4는 일반화 선형모형의 각 내재적 모형에서의 차수 추정 결과를 상위 3개의 차수에 대한 퍼센트 막대그래프로 나타낸 것이다. 단, x축은 추정된 차수를 나타내며, N에 관계없이 총 4,000번의 실험을 요약하였다. past_obs 차수만 있는 경우 추정의 성능이 아주 높지만, past_mean의 차수가 추가되거나 높아질수록 추정의 정확도는 떨어진다.

추가적으로 일반화 선형모형의 분포와 연결함수를 다르게 설정한 경우에도 ARIMA 모형에 비해 좋은 예측 성능을 갖는지 살펴보고자 한다. 2가지 분포(포아송, 음이항)와 2가지 연결함수(항등, 로그)의 조합에 관계없이 모두 Table 3.1과 유사한 결과(지면 관계상 생략)를 보였으나, 음이항 분포에서 연결함수가 로그함수인 경우, Table 3.2의 결과와 같이 일반화 선형 모형의 성능은 $N = 50$인 경우를 제외하고는 ARIMA 모형에 비해서 좋지 않았으며 종종 RMSE 및 MAE의 값이 상대적으로 아주 크게 나타나기도 하였다. $N = 100$ 이상에서 ARIMA 모형의 성능이 특히 개선된 것은, 과산포의 특징을 가지는 음이항 분포에서 로그 연결함수가 분산안정화 변환의 역할을 함으로써 생긴 결과로 추측된다.

3.2.2. 내재적 모형이 ARMA 모형일 경우 내재적 모형이 ARMA 모형인 경우 ARMA$(1, 0)$, ARMA$(2, 0)$, ARMA$(3, 0)$, ARMA$(1, 1)$을 고려하여 3.2.1절과 동일한 방식으로 반복 실험하였으며 결과는 Table 3.3과 같다. 본 논문은 모형 (2.1)을 모수 공간 (2.6) 또는 (2.7)과 같은 정상성 조건 하에서 고려하기 때문에 차분이 있는 ARIMA 모형이 내재적 모형인 경우는 예측 성능 검토를 위한 실험을 생략하기로 한다.

시계열 길이가 50, 100인 상황을 제외한 대부분의 상황에서 일반화 선형 모형의 평균 RMSE와 평균 MAE의 값이 ARIMA 모형에 비해 더 작게 나타났다. 따라서 내재적 모형이 비계절 ARMA이고 차수

Table 3.2. Simulation results in case of the negative binomial distribution with the log link

True model	$N = 50$		$N = 100$		$N = 200$		$N = 400$	
	ARIMA	TSGLM	ARIMA	TSGLM	ARIMA	TSGLM	ARIMA	TSGLM
				mean_RMSE				
TSGLM(1)	115.84	**110.77**	**115.65**	162.06	**123.79**	130.87	**125.83**	1.75e+70
TSGLM(2)	**105.85**	268.15	**113.15**	113.74	**128.62**	134.10	**131.99**	138.77
TSGLM(3)	**99.02**	106.31	**105.88**	345.74	**120.50**	597.99	**126.33**	135.49
TSGLM(1)(1)	90.56	**89.27**	**91.58**	93.01	**94.23**	99.89	**95.28**	103.76
TSGLM(2)(1)	**83.05**	83.79	**88.22**	156.35	**87.85**	91.11	**88.90**	96.76
TSGLM(1)(2)	84.60	**82.06**	**84.36**	86.42	**88.36**	91.69	**87.56**	96.11
				mean_MAE				
TSGLM(1)	94.97	**89.17**	**89.47**	106.36	**90.22**	102.82	**83.97**	1.96e+69
TSGLM(2)	**82.47**	137.17	**84.19**	84.30	**91.55**	98.30	**88.36**	103.21
TSGLM(3)	75.65	**75.18**	**76.15**	149.02	**83.53**	224.89	**82.89**	99.22
TSGLM(1)(1)	73.62	**72.432**	**72.611**	75.46	**71.311**	81.059	**69.514**	84.082
TSGLM(2)(1)	67.29	**66.67**	**69.87**	88.55	**67.17**	73.48	**66.07**	79.37
TSGLM(1)(2)	68.64	**66.42**	**66.25**	69.12	**67.46**	74.07	**64.84**	78.58

Table 3.3. Simulation result of RMSE, MAE for the two models: ARMA as true model

True model	$N = 50$		$N = 100$		$N = 200$		$N = 400$	
	ARIMA	TSGLM	ARIMA	TSGLM	ARIMA	TSGLM	ARIMA	TSGLM
				mean_RMSE				
ARMA(1, 0)	1.5985	**1.4966**	1.7194	**1.5779**	1.7832	**1.6387**	1.7288	**1.6557**
ARMA(2, 0)	1.3832	**1.3712**	1.4675	**1.3785**	1.5828	**1.4457**	1.5551	**1.4459**
ARMA(3, 0)	**1.3220**	1.3833	**1.5304**	1.5478	1.7344	**1.6248**	1.9181	**1.7042**
ARMA(1, 1)	1.3595	**1.2875**	1.3644	**1.3187**	1.3318	**1.3116**	1.3458	**1.3351**
				mean_MAE				
ARMA(1, 0)	1.3742	**1.2797**	1.4489	**1.3184**	1.4759	**1.3413**	1.4038	**1.3367**
ARMA(2, 0)	1.1849	**1.1750**	1.2313	**1.1474**	1.3136	**1.1847**	1.2743	**1.1726**
ARMA(3, 0)	**1.1265**	1.1901	**1.2935**	1.3152	1.4600	**1.3588**	1.6058	**1.4063**
ARMA(1, 1)	1.1416	**1.0756**	1.1195	**1.0772**	1.0771	**1.0576**	1.0826	**1.0728**

가 비교적 단순한 경우에 시계열 일반화 선형 모형의 예측 성능이 ARIMA 모형의 예측 성능보다 우수한 것을 확인할 수 있다.

3.2.3. 내재적 시계열의 평균에 따른 시뮬레이션 일반적으로 관측치의 값이 큰 경우, 중심극한정리(central limit theorem)에 의하여 경험적 분포의 형태는 대칭적이며 정규 분포에 가까워지는 경향이 있다. 즉, 시계열의 평균이 큰 경우 계수형 시계열 자료 분석에 ARIMA 모형도 사용 가능하며 이 경우 일반화 선형 모형이 ARIMA 모형에 비해 좋은 성능을 갖는지 살펴보고자 한다. 이를 위하여 내재적 프로세스(underlying or true process)의 평균의 크기를 $\mu = 10, 40, 100, 200$으로 다르게 설정하여 각각 $N = 100$개의 관측치를 생성하였다. 내재적 모형은 ARMA$(1, 0)$, ARMA$(2, 0)$, ARMA$(0, 1)$, ARMA$(0, 2)$, ARMA$(1, 1)$으로 각각 설정하였으며, 3.2.1절과 동일한 방식으로 실험하였으며 결과는 Table 3.4를 통해 확인할 수 있다.

그러나, 기대하였던 ARIMA 모형의 성능은 $\mu = 100$일 때 ARMA$(1, 0)$의 경우에서만 우수하였으며 나머지 모든 상화에서는 일반화 선형 모형의 평균 RMSE와 평균 MAE의 값이 더 작게 나타났다. 따라서

Table 3.4. Simulation result of RMSE, MAE for the two models according to means

True model	$\mu = 10$		$\mu = 40$		$\mu = 100$		$\mu = 200$	
	ARIMA	TSGLM	ARIMA	TSGLM	ARIMA	TSGLM	ARIMA	TSGLM
mean_RMSE								
ARMA(1, 0)	1.1712	**1.1512**	1.1481	**1.1237**	1.1629	**1.1398**	1.2775	**1.2328**
ARMA(2, 0)	1.4674	**1.4175**	1.4810	**1.3821**	1.4902	**1.4013**	1.5161	**1.4186**
ARMA(0, 1)	1.1110	**1.1004**	1.1220	**1.1116**	1.1229	**1.1084**	1.1283	**1.1156**
ARMA(0, 2)	1.1461	**1.1406**	1.1519	**1.1446**	1.1559	**1.1487**	1.1456	**1.1362**
ARMA(1, 1)	1.5302	**1.4784**	1.5181	**1.4936**	1.5522	**1.4982**	1.5588	**1.5144**
mean_MAE								
ARMA(1, 0)	0.9570	**0.9388**	0.9392	**0.9172**	**0.8016**	0.9293	1.0501	**1.0063**
ARMA(2, 0)	1.2307	**1.1835**	1.2436	**1.1523**	1.2501	**1.1686**	1.2797	**1.1881**
ARMA(0, 1)	0.9022	**0.8913**	0.9096	**0.8987**	0.9109	**0.8971**	0.9161	**0.9036**
ARMA(0, 2)	0.9350	**0.9301**	0.9355	**0.9294**	0.9404	**0.9342**	0.9325	**0.9243**
ARMA(1, 1)	1.2627	**1.2116**	1.2469	**1.2231**	1.2784	**1.2279**	1.2781	**1.2367**

평균의 크기에 상관없이 내재적 모형이 비계절 ARMA인 경우에도 ARIMA 모형의 예측 성능보다 시계열 일반화 선형 모형의 성능이 더 우수한 것을 확인할 수 있다.

4. 실증분석

4.1. 분석 자료

본 장에서는 시계열 일반화 선형 모형을 사용하여 실제 계수형 시계열 자료를 적합 및 예측한다. 이를 통하여 본 연구에서 제안하는 계수형 시계열 모형의 자동화 차수 결정 알고리즘을 검증한다. 실증분석에 사용한 자료는 검찰청에 보고된 국내 살인사건 발생 건수이며, 2002년 1월부터 2013년 12월까지의 월별 자료이다. 총 144개의 관측값으로 구성된 계수형 시계열 자료로 모형 적합을 위한 훈련자료(training set)의 기간은 2002년 1월부터 2010년 12월이고, 예측 성능 평가를 위한 검증 자료(test set)의 기간은 2011년 1월부터 2013년 12월이다. 해당 자료는 국가통계포털을 이용하여 얻을 수 있다 (http://kosis.kr).

국내의 살인사건 발생 건수에 대한 시계열 그림은 Figure 4.1과 같다. 2002년 1월부터 2010년 7월까지 증가하다가 이후로는 감소하는 추세를 보인다. 계절성이 있는지 확인하기 위해 월별 발생 건수를 1년 단위로 그리면 Figure 4.2와 같다. 연별 살인사건 발생 건수를 비교했을 때 전반적으로 7월까지는 증가하다가 이후로 다시 감소하는 패턴을 보인다. 따라서 매년 2월에 살인사건 발생 건수가 최소이고, 대체로 7월에 살인사건 발생 건수가 최대인 일정한 패턴이 반복되기 때문에 12개월의 계절 주기를 갖는다고 할 수 있다.

4.2. 예측 성능 비교

예측 성능을 비교하기 위한 대조 모형으로 ARIMA 모형을 사용하였으며, 두 모형의 예측 성능을 비교하는 지표는 RMSE, MAE 그리고 mean absolute percentage error (MAPE)를 사용하였다. 실제 시계열 자료 y_i와 예측값 $\hat{y}_i$에 대하여 미래 시점 h까지의 MAPE는 다음과 같이 계산한다.

$$\text{MAPE} = \frac{100}{h}\sum_{i=1}^{h}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{4.1}$$

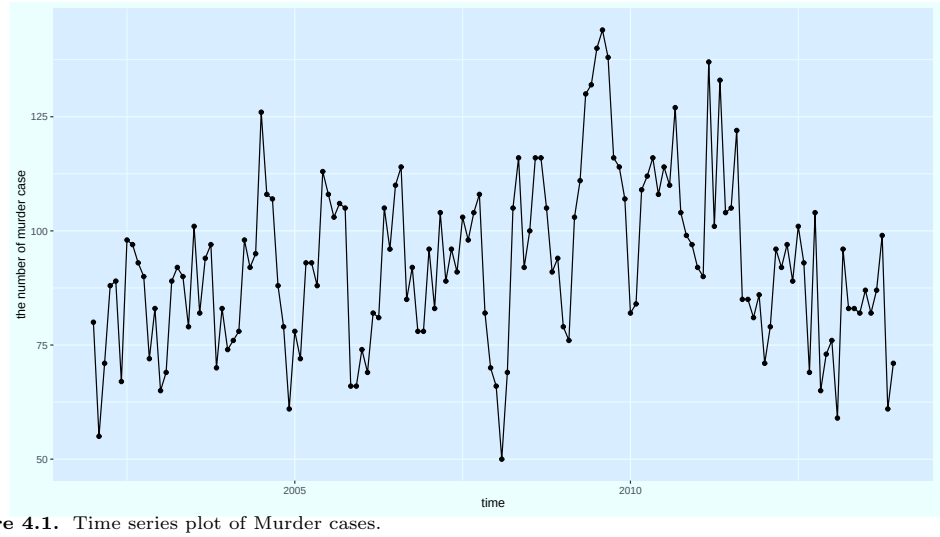

Figure 4.1. Time series plot of Murder cases.

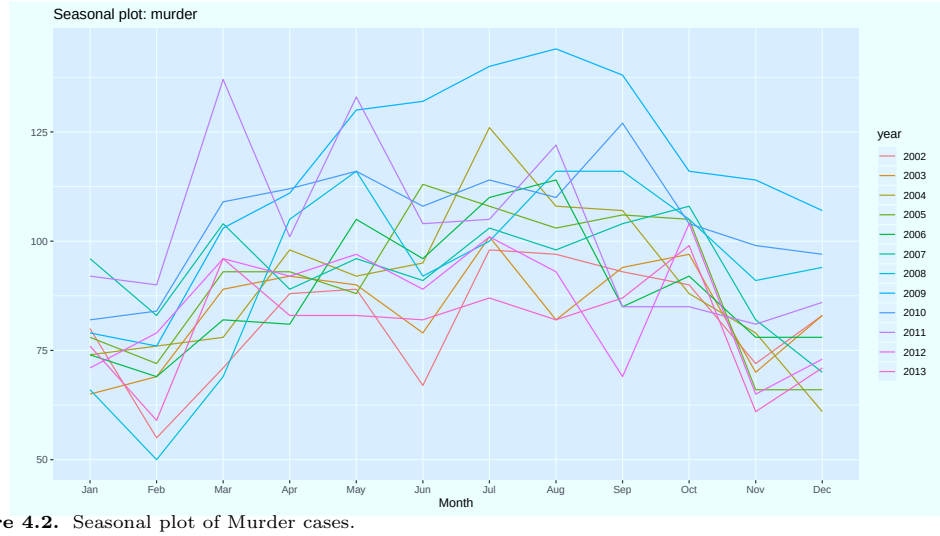

Figure 4.2. Seasonal plot of Murder cases.

살인 사건 발생 건수 자료에서 auto.arima 함수를 이용하여 ARIMA 모형의 차수를 추정하면 ARIMA $(1,0,0)(2,1,0)_{s=12}$이 선택된다. 또한, 본 논문에서 제안한 계수형 시계열 모형을 위한 자동화 알고리즘을 적용한 결과 시계열 일반화 선형 모형은 TSGLM$(1,11,12,24)(0)$이 선택되었다.

Figure 4.3은 최종 선택된 두 모형을 이용하여 예측한 36개월의 예측값과 검증 자료에 대한 그림이다. 검증 자료 기간 동안 매년 감소하는 추세를 보이는 실제 살인 사건 발생 건수와는 대조적으로, ARIMA 모형의 경우 증가하는 추세를 예측하였으며 예측값 자체도 실제값에서 많이 벗어난다. 반면, 시계열 일반화 선형 모형의 경우 2011년 1월부터 2013년 12월까지의 실제 살인 사건 발생 건수의 흐름과 전반적으로 비슷한 추세를 예측하였다. 또한, Table 4.1과 같이 3개의 예측 성능 비교 지표 RMSE, MAE, MAPE의 모든 값에서 시계열 일반화 선형 모형이 ARIMA에 비해 작은 값을 갖는 것을 확인할 수 있

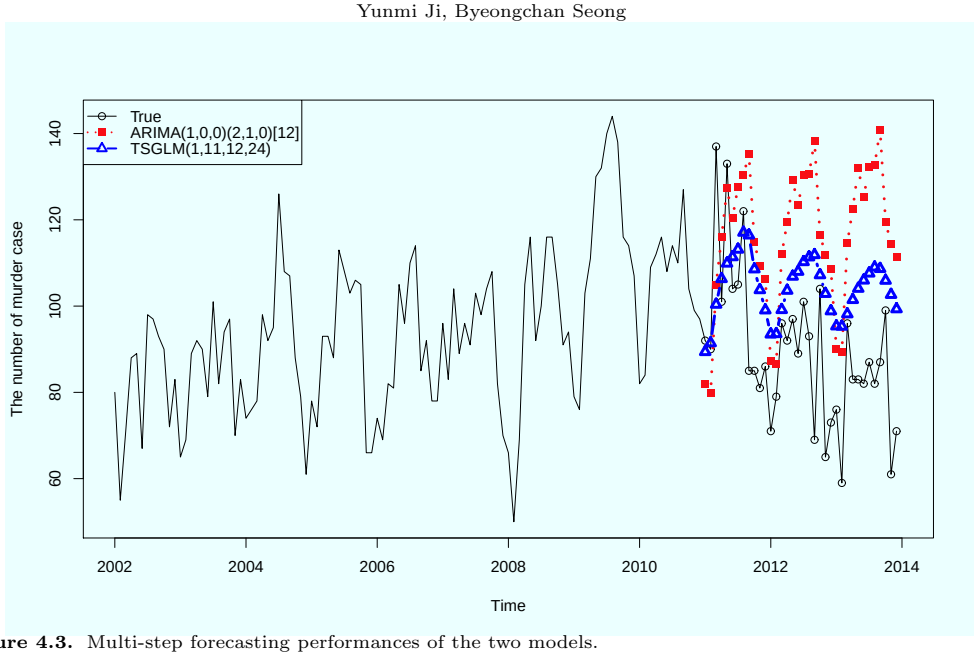

Figure 4.3. Multi-step forecasting performances of the two models.

Table 4.1. RMSE, MAE, MAPE for test set forecasts of the two models

Model	RMSE	MAE	MAPE
$\text{ARIMA}(1,0,0)(2,1,0)_{12}$	33.72662	29.81188	36.2572
$\text{TSGLM}(1,11,12,24)$	**21.88381**	**18.49248**	**23.0145**

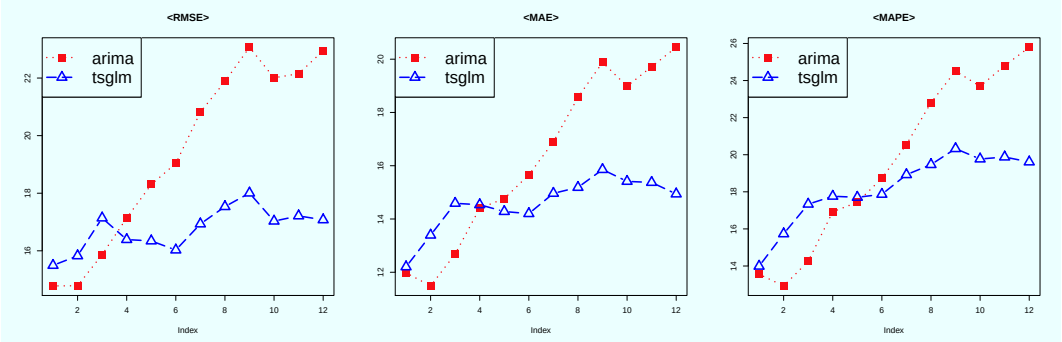

Figure 4.4. Comparison of forecast performance of the two models applied Murder cases.

다. 따라서 Figure 4.3과 Table 4.1를 통해 해당 자료에서 ARIMA모형에 비해 일반화 선형 모형의 예측 성능이 더욱 우수함을 알 수 있다.

4.3. 시계열 교차검증

추가적으로 시계열 교차검증(cross-validation) 방법을 사용하여 시계열 일반화 선형 모형의 예측 성능을 평가하였다. 최초의 훈련 자료 기간은 2002년 1월부터 2009년 12월이고, 2010년 이후의 시점을 월별로 하나씩 추가하여 훈련 자료를 갱신하였다. 각 훈련 자료에서 최대 12시점까지 예측하였으며, 동일 미래 시점 예측에 대하여 Figure 4.4와 같이 예측 성능을 비교하였다. ARIMA와 시계열 일반화 선

형 모형의 예측 성능을 비교했을 때, 1, 2, 3시점 이후의 예측에서는 ARIMA 모형의 RMSE, MAE, MAPE 값이 일반화 선형 모형의 경우보다 더 작은 것을 확인할 수 있다. 이는 살인 사건 발생 건수 자료의 경우 계절성이 뚜렷하고 차수가 비교적 복잡하기 때문에 단기 예측에서 ARIMA 모형의 성능이 더 우수한 것으로 보여진다. 6시점 이상의 예측에서는 시계열 일반화 선형 모형의 RMSE, MAE, MAPE 값이 더 작게 나타나며 중·장기 예측에서는 시계열 일반화 선형 모형이 ARIMA 모형에 비해 더 좋은 예측 성능을 갖는 것을 확인할 수 있다. 또한, ARIMA의 경우에는 RMSE, MAE, MAPE가 예측 시점이 멀어질수록 급격히 증가하지만 시계열 일반화 선형 모형은 전반적으로 비슷한 값을 갖는 것을 확인할 수 있다. 즉, ARIMA 모형은 예측기간이 증가할수록 예측력이 급격하게 감소하지만 시계열 일반화 선형 모형은 미래 예측 시점에 상관없이 안정적인 예측력을 보인다.

5. 결론

본 논문에서는 계수형 시계열 자료에 대하여 포아송 또는 음이항 분포를 가정하는 시계열 일반화 선형 모형에 대해 소개하고 모형에 사용되는 차수를 자동으로 결정하는 자동화 알고리즘을 제안하고 있다. 또한, 제안한 자동화 차수 결정 알고리즘을 이용하여 시뮬레이션을 통한 시계열 일반화 선형 모형과 ARIMA 모형의 예측 성능을 비교하고, 실증분석으로 국내 살인 사건 발생 건수에 대하여 분석하였다. 시뮬레이션 및 실증분석 결과, 차수가 비교적 단순한 경우 시계열 일반화 선형 모형의 예측 성능이 ARIMA 모형의 예측 성능보다 우수한 것으로 나타났다. 반면, 계절성이 있고 자기상관 구조가 복잡한 경우 단기 예측에서는 ARIMA 모형의 예측 성능이 더 우수하였으나 예측 시점이 멀어질수록 시계열 일반화 선형 모형의 예측 성능이 더 안정적으로 나타났다.

References

Christou, V. and Fokianos, K. (2014). Quasi-likelihood inference for negative binomial time series models, *Journal of Time Series Analysis*, **35**, 55–78.

Doukhan, P., Fokianos, K., and Tjøstheim, D. (2012). On weak dependence conditions for Poisson autoregressions, *Statistics & Probability Letters*, **82**, 942–948.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models* (2nd ed), Springer, New York.

Ferland, R., Latour, A., and Oraichi, D. (2006). Integer-valued GARCH process, *Journal of Time Series Analysis*, **27**, 923–942.

Fokianos, K., Rahbek, A., and Tjøstheim, D. (2009). Poisson autoregression, *Journal of the American Statistical Association*, **104**, 1430–1439.

Fokianos, K. and Tjøstheim, D. (2011). Log-linear Poisson autoregression, *Journal of Multivariate Analysis*, **102**, 563–578.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R, *Monash Econometrics and Business Statistics Working Papers* 6/07, Monash University, Department of Econometrics and Business Statistics.

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

Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis*, John Wiley & Sons, Chichester.

Liboschik, T., Fokianos, K., and Fried, R. (2017). tscount: An R package for analysis of count time series following generalized linear models, *Journal of Statistical Software*, **82**, 1–51.

Tjøstheim, D. (2015). Count time series with observation-driven autoregressive parameter dynamics, *Handbook of Discrete-Valued Time Series, Handbooks of Modern Statistical Methods*, 77–100.

Weiß, C. H. (2008). Thinning operations for modeling time series of counts—a survey, *AStA Advances in Statistical Analysis*, **92**, 319.

계수형 시계열 모형을 위한 자동화 차수 선택 알고리즘

지윤미[a] · 성병찬[a,1]

[a]중앙대학교 응용통계학과

(2019년 12월 30일 접수, 2020년 1월 12일 수정, 2020년 1월 12일 채택)

요 약

본 논문은 시계열 일반화 선형 모형의 하나인 계수형 시계열 모형에서 중요한 역할을 하는 과거 관측값과 조건부 평균값의 차수를 자동으로 결정하는 알고리즘을 연구한다. 본 알고리즘은 ARIMA 모형의 차수를 기반으로 시계열 일반화 선형 모형의 차수 후보군을 만들고, 차수 후보군의 조합을 이용하여 정보량 기준으로 최종 모형으로 선택한다. 제안된 알고리즘을 평가하기 위하여, 내재적 모형 및 내재적 시계열의 종류에 따른 시뮬레이션 및 실증 분석을 수행하고 예측력을 ARIMA 모형과 비교한다. 예측 성능 평가 결과, 계수형 시계열 분석에서 ARIMA 모형에 비해 시계열 일반화 선형 모형의 예측 성능이 우수함을 확인할 수 있다. 또한 실증분석으로서, 살인사건 발생 건수의 예측 결과 ARIMA 모형보다 중기 및 장기 예측에서 우수한 성능을 나타내는 것을 확인할 수 있다.

주요용어: 계수형 시계열, 자동화 알고리즘, 시계열 일반화 선형 모형, ARIMA 모형

이 논문은 2018년도 중앙대학교 연구장학기금 지원에 의한 것이며, 제1저자 지윤미의 석사학위논문을 수정하여 작성한 것임.

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