The Korean Journal of Applied Statistics
2023, Vol. 36, No. 4, 309–322

DOI: https://doi.org/10.5351/KJAS.2023.36.4.309
Print ISSN1225-066X / Online ISSN 2383-5818

An overview of Hawkes processes and their applications

Mijeong Kim[1,a]

[a]Department of Statistics, Ewha Womans University

Abstract

The Hawkes process is a point process with self-exciting characteristics. It has been mainly used to describe seismic phenomena in which aftershocks occur due to the main earthquake. Recently, it has been used to explain various phenomena with self-exciting properties, such as the spread of infectious diseases and the spread of news on SNS. The Hawkes process can be flexibly modified according to the characteristics of events by using various types of excitation functions. Since it is difficult to implement a maximum likelihood estimator numerically, estimation methods have been improved until recently. In this paper, the conditional intensity function and excitation function are explained to describe the Hawkes process. Then, existing examples of Hawkes processes used in seismic, epidemiological, criminal, and financial fields are described and estimation methods are introduced. I analyze earthquakes that occurred in gyeongsang-do, Korea from November 2017 to December 2022, using R package `ETAS`.

Keywords: conditional intensity function, ETAS, excitation function, Hawkes process, self-exciting

1. 서론

혹스 과정(Hawkes process)은 자기 자극 특성을 가진 확률 과정(stochastic process) 모형으로 Hawkes (1971)에 처음 소개되었다. 자기 자극 특성은 지진을 예로 들어 간단히 설명할 수 있다, 지진이 한번 발생하면 수많은 여진이 연이어서 발생한다. 지진이라는 한 사건이 다른 많은 사건, 즉 여진을 발생시키는데, 이것을 자기 자극(self-exciting)이라고 한다. 혹스 과정은 지진의 발생을 모형화하는데 주로 사용되었는데, 특히 Ogata (1988)가 제시한 epidemic-type aftershock sequence (ETAS) 모형은 여진 발생을 설명하기 위해 수정된 Omori의 법칙을 이용한 혹스 모형으로서 지진학에서 지금까지도 중요한 모형으로 활용되고 있다. 그 이후에도 다양한 자극 함수의 도입으로 ETAS 모형이 개선되고 발전되었다 (Ogata와 Katsura, 1998; Ogata, 1998; Zhuang 등, 2002; Daley와 Vere-Jones, 2008; Mohler 등, 2011). Hawkes (2018)에 따르면, 혹스 과정이 소개된 후 25년 가량은 지진학 이외의 분야에서는 거의 활용되지 않다가 2000년 초반부터 금융, 범죄학 및 역학 분야 등 다양한 분야에 활용되기 시작하였다. 최근 들어, 혹스 과정 모형은 소셜 네크워크를 통한 뉴스 및 영상의 전파를 설명할 수 있는 유용한 모형으로 여기지고 있으며, 최근에 유행한 COVID19을 모형화하는데도 활용되기도 하였다.

자기 자극 특성을 설명하기 위해 도입된 함수를 자극 함수(excitation function)이라고 하는데, 이 자극 함수를 다양하게 변형시킴으로써 다양한 형태의 혹스 과정 모형을 만들 수 있다. 분석하고자 하는 사건의 특성에 따라 지수붕괴 함수 및 멱법칙 함수와 같은 단순한 형태 뿐만 아니라 비교적 복잡한 isotropic kernel 등

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government (NRF-2020R1F1A1A01074157).

[1] Department of Statistics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail : m.kim@ewha.ac.kr

Published 31 August 2023 / journal homepage: http://kjas.or.kr

다양한 형태의 자극 함수를 선택할 수 있다. 모형이 복잡해질수록 모수 추정은 어려운 작업이 되기 때문에, 최근까지도 개선된 추정 방법에 대한 연구가 계속 되고 있다.

이 논문은 다음과 같이 구성되어 있다. 2장에서는 계수 과정과 점 과정을 정의하고, 조건부 강도 함수를 설명한 후 혹스 과정을 정의한다. 3장에서는 몇 가지 대표적인 혹스 과정의 종류에 대한 설명을 하고, 4장에서는 혹스 과정이 적용된 예 중심으로 설명하고자 한다. 5장에서는 혹스 과정의 추정 방법에 대해 설명하고, 6장에서는 ETAS 모형을 이용하여 경상도에서 2017년 11월부터 2022년 12월까지 발생한 지진을 분석한다. 7장에서는 이 논문을 요약한다.

2. 혹스 과정

2.1. 계수 과정과 점 과정

계수 과정(counting process)과 점 과정(point process)의 의미를 살펴 보자. 계수 과정 $\{N(t), t \geq 0\}$는 확률 과정(stochastic process)으로서 비음(nonnegative)인 정수 값을 갖고, $t = 0$일 때에는 $N(0) = 0$이다. $N(t)$는 주로 시간 t까지 일어난 사건의 수로 설명된다. $s < t$일 때, $N(t) - N(s)$는 $(s, t]$ 구간에서 일어난 사건의 수가 된다. 점 과정은 확률 변수(random variable)의 수열 $T = \{T_1, T_2, \ldots\}$로 표현되는데, 집합 T의 원소 $T_1, T_2, \ldots$는 $[0, \infty)$ 내에서 값을 갖고, $P(0 \leq T_1 \leq T_2 \leq \ldots) = 1$이다. 일정한 시간 안에서 일어나는 사건의 수를 계수 과정으로 볼 수 있고, 사건이 발생 시점들의 집합은 점 과정으로 볼 수 있다. 계수 과정의 대표적인 예로 포아송 과정(Poisson process)을 들 수 있다.

점 과정은 과거에 일어난 사건을 조건부로 하여 다음 도착 시점을 분포 함수로 하여 수학적으로 표현할 수 있다. 가장 최근에 사건이 일어난 시점이 u이고 누적된 사건의 수는 k, u 시점까지의 정보(history)를 $\mathcal{H}(u)$라고 할 때, 다음 도착 시점 T_{k+1}의 조건부 누적 확률 분포(conditional cdf)는 다음과 같다.

$$F^*(t \mid \mathcal{H}(u)) = \int_u^t P(T_{k+1} \in [s, s+ds] \mid \mathcal{H}(u))\, ds = \int_u^t f^*(s \mid \mathcal{H}(u))\, ds.$$

실현된 값 $\{t_1, t_2, \ldots, t_k\}$의 결합 확률 밀도 함수는 연쇄 법칙(chain rule)을 이용하여 다음과 같이 구할 수 있다.

$$f(t_1, t_2, \ldots, t_k) = \prod_{i=1}^{k} f^*(t_i \mid \mathcal{H}(t_{i-1}))\,.$$

사건이 일어났다는 것은 영어권에서는 사건이 도달했다는 의미로 arrival이라는 단어를 쓴다. $f^*(t|\mathcal{H}(u))$를 조건부 도착 분포(conditional arrrival distribution)라고 하며, 간단히 $f^*(t)$로 표시하겠다.

2.2. 조건부 강도 함수

점 과정을 조건부 도착 분포 $f^*(t)$로 설명하는 것이 어렵기 때문에, 짧은 시간 동안에 사건이 일어나는 비율을 뜻하는 강도(intensity)를 이용한 조건부 강도 함수(conditional intensity function)를 도입하였다. 조건부 강도 함수는 생존 분석에서 주로 이용하는 위험 함수(hazard function)과 같다. 참고로 위험 함수는 다음과 같다.

$$\lambda^*(t) = \frac{f^*(t)}{1 - F^*(t)}.$$

계수 과정에서는 조건부 강도 함수를 다음과 같이 정의한다.

Definition 1. 조건부 강도 함수(conditional intensity function) *계수 과정 $N(\cdot)$에 해당하는 과거 사건들에 대한 누적된 정보(history)에 대한 함수를 $\mathcal{H}(\cdot)$라고 하자.*

$$\lambda^*(t) = \lim_{h\to 0} \frac{E\left[N(t+h) - N(t) \mid \mathcal{H}(t)\right]}{h}. \tag{2.1}$$

위와 같이 비음의 함수 $\lambda^*(t)$가 존재하면, $\lambda^*(t)$를 조건부 강도 함수라고 한다.

혹스 과정을 정의하기 전에 이해를 돕기 위해 포아송 과정을 먼저 설명하고자 한다.

Definition 2. 포아송 과정(Poisson process) *계수 과정 $\{N(t), t \geq 0\}$가 다음을 만족하면 포아송 과정이라고 한다.*

1. $N(0) = 0$.
2. $s < t$*에 대하여* $N(s)$*와* $N(t) - N(s)$*는 독립이다.*
3. $N(t)$*와* $N(s+t) - N(s)$*는 평균이* νt*인 동일한 포아송 분포를 갖는다. 즉*

$$P(N(t) = k) = P(N(s+t) - N(s) = k) = e^{-\nu t}\frac{(\nu t)^k}{k!}, \quad s > 0,\ k \geq 0.$$

포아송 과정을 사건이 일어나는 시점에 초점을 맞추어 설명할 때에는 포아송 점 과정(Poisson point process)이라고 한다. 포아송 과정의 조건부 강도는 다음과 같이 사건이 일어나는 시점에 상관 없이 ν로 일정하다.

$$\lambda^*(t) = \lim_{h\to 0} \frac{E\left[N(t+h) - N(t) \mid \mathcal{H}(t)\right]}{h} = \lim_{h\to 0} \frac{E(N(h))}{h} = \nu.$$

즉, 특정 시간 간격(time interval) 동안 일어나는 사건의 수는 시간 간격이 겹치지 않는다면 이전에 일어난 사건의 수와 독립이고 그 비율이 일정한데, 이러한 특성을 비기억성(memoryless property)이라고 한다. 백화점에 들어오는 고객의 수, 지진 발생 건수, 질병에 확진된 사람들의 수 등과 같이 일상 생활에서 계수 과정으로 설명할 수 있는 현상이 많이 존재한다. 많은 경우에 앞서 일어난 사건의 수는 다음 일어난 사건의 수와 관련이 있거나, 사건이 일어난 시점에 따라 사건이 일어난 비율이 다를 수 있다. 예를 들어 백화점에 들어 오는 고객의 수는 시간 간격이 겹치지 않으면 과거 시간에 들어 온 고객의 수와 독립으로 볼 수 있으나, 오전에 입장하는 고객의 비율과 오후에 입장하는 고객의 비율이 다를 수 있다. 이런 경우에는 조건부 강도 함수를 시점 별로 다르게 변형한 비동질적 포아송 과정(nonhomogeneous Poisson process)을 이용할 수 있다. 비동질적 포아송 과정과 구별하기 위하여 위에서 정의한 포아송 과정을 동질적 포아송 과정(homogeneous Poisson process)라고 한다. 비동질적 포아송 과정은 강도 함수 $\lambda^*(t)$가 상수가 아니라 시간의 함수라는 점만 제외하고 포아송 과정의 모든 속성을 따른다. 비동질적 포아송 과정에서는 다음 식이 성립한다.

$$N(t+s) - N(t) \sim \text{Poisson}\left(\int_t^{t+s} \lambda^*(u)\,du\right).$$

확률 과정이 시간 이동에 따라 그 분포가 변하지 않으면(invariant under time shift), 시간 정상성(stationarity in time)이 있다고 정의한다. 동질적 포아송 과정은 시간 정상성을 만족하고 비동질적 포아송 과정은 시간 비정상성(non-stationarity) 특성을 갖는다. 지진의 발생과 전염병 감염의 경우는 시간 간격이 겹치지 않더라도 앞서 일어난 사건의 수와 다음에 일어난 사건의 수는 밀접한 관련이 있다. 이러한 경우에는 조건부 강도 함수를 좀 더 유연한 형태로 만들면 현상을 수학적으로 설명하기가 용이해진다. 조건부 강도 함수는 주로 자기 자극(self-exciting)과 자기 조절(self-regulating)라는 용어로 설명된다. 자기 자극은 이전 시간 간격에 발생한 사건이 미래에 발생할 사건의 가능성을 높이는 것을 의미하고, 반대로 자기 조절은 이전 시간 간격에 발생한 사건이 미래에 발생할 사건의 가능성을 낮추는 것을 의미한다.

2.3. 혹스 과정

혹스 과정은 자기 자극 또는 자기 조절의 특성을 갖는 계수 과정으로서 다음과 같이 정의된다.

Definition 3. 혹스 과정(Hawkes process) *계수 과정 $N(\cdot)$에 해당하는 과거 사건들에 대한 누적된 정보에 대한 함수를 $\mathcal{H}(\cdot)$라고 할 때, 계수 과정 $\{N(t), t \geq 0\}$가 다음을 만족하면 혹스 과정이라고 한다.*

$$P\left(N\left(t+h\right)-N\left(t\right)=n \mid \mathcal{H}\left(t\right)\right)=\begin{cases} \lambda^{*}\left(t\right)h+o\left(h\right), & n=1, \\ o\left(h\right), & n>1, \\ 1-\lambda^{*}\left(t\right)h+o\left(h\right), & n=0, \end{cases}$$

여기서 $o(h)$는 $h \to 0$일 때, h에 비해 더 빨리 0으로 근접해가는 항을 나타낸다. 만약 $a(h) = o(h)$라면, $\lim_{h\to 0} a(h)/h = 0$이다. 혹스 과정의 조건부 강도 함수는 다음과 같다.

$$\lambda^{*}\left(t\right)=\mu+\int_{0}^{t} g\left(t-u\right)dN\left(u\right). \tag{2.2}$$

이 때 상수 $\mu > 0$를 배경 자극(background intensity)이라고 하고, $g : (0, \infty) \to [0, \infty)$는 시간에 따른 자기 자극의 정도를 반영한 함수로서, 자극 함수(excitation function) 또는 자극 커널(exciting kernel 또는 triggering kernel)이라고 한다. μ 뒤에 따라오는 다른 항이 없다면 포아송 과정이 되므로 자기 자극 또는 자기 조절의 영향을 받지 않는 포아송 기본 수준(Poisson base level)이라고 한다. 혹스 과정의 조건부 강도 함수는 배경 자극과 자극 함수를 이용하여 유연한 형태로 만들 수 있다. 자극 함수에는 t 시점 이전에 발생하는 사건이 시점 t에 기여하는 정도가 반영되어 있다. 자극 함수 $g(\cdot) = 0$이면 포아송 분포가 되므로, 혹스 과정에서는 $g(\cdot) \neq 0$을 가정한다. 자극 함수 $g(\cdot)$에 대해서는 다음 장에서 자세히 설명할 것이다. 혹스 과정은 모형에 따라 시간 정상성을 만족하기도 하고 시간 비정상 특성을 갖기도 한다. t 시점까지 실현된 사건의 시점의 집합을 $\{t_1, t_2, \ldots, t_k\}$라고 하면, 혹스 조건부 강도(Hawkes conditional intensity) (2.2)는 다음과 같이 표현할 수 있다.

$$\lambda^{*}\left(t\right)=\mu+\sum_{t_i<t} g\left(t-t_i\right). \tag{2.3}$$

3. 혹스 과정의 종류

앞 장에서도 언급하였듯이, 혹스 과정의 조건부 강도 함수는 배경 자극과 자극 함수로 결정된다. 자극 함수 선택을 통해 혹스 과정 모형을 다양하게 만들 수 있다.

3.1. 일변량 혹스 과정

1. 분기 과정(branching process)

   Hawkes와 Oakes (1974)는 분기 과정을 제안하였다. 분기는 나뉘어서 갈라진다는 뜻으로, 분기 과정은 T_i 시점에 발생한 하나의 사건을 아버지나 어머니(a parent)로 볼 수 있고, 부모로부터 자녀가 태어나는 것처럼 이 사건으로부터 다른 사건들이 발생하는 점 과정을 일컫는다. 한 사건으로부터 다른 사건이 야기되는 비율(자손을 낳는 비율)을 (2.3)의 자극 함수 $g(t - T_i), t > T_i$에 반영하여 조건부 강도 함수를 만들 수 있다. 분기 과정은 바이러스의 전파, 생명을 가진 개체 집단의 번식을 설명하는 모형에 활용될 수 있다.

2. 지수 붕괴(exponential decay) 함수의 사용

간단한 형태의 자극 함수로서 다음과 같은 형태를 갖는다.

$$g(u) = \alpha e^{-\beta u}, \quad u > 0. \tag{3.1}$$

이 때, $\alpha, \beta > 0$이고, α는 전반적인 강도(overall strength of the excitation), 비율 상수(rate constant) β는 과거 사건으로부터 영향받아 일어나는 다음 사건의 속도를 조절한다. 새로운 사건이 일어날 때마다 조건부 강도는 α만큼 증가한 다음, 시간이 지남에 따라 사건 발생에 대한 영향은 β 비율로 지수적으로 감소한다.

3. 멱법칙(Powerlaw) 함수의 사용

$$g(u) = \frac{k}{(c+u)^p}, \quad u > 0.$$

이 때, $c, k, p > 0$이다. 멱법칙 함수를 이용한 혹스 과정은 지진과 금융 분야에서 장기적인 영향을 설명하는 데 자주 사용된다.

4. 일반화된 형태의 혹스 과정

$$\eta(\lambda^*(t)) = \mu(t) + \sum_{t_i < t} g(t - t_i, m_i).$$

조건부 강도에 영향을 줄 수 있는 mark m_i가 추가되어 좀 더 유연한 모형을 만들 수 있다. 예를 들어, 금융 시장에서 거래량이 많은 거래는 거래량이 적은 거래보다 미래의 거래에 더 크게 영향을 줄 수 있으므로, 거래량을 mark로 모형에 포함시키면 이런 차이를 반영하는 데 도움이 된다. 이러한 확률과정을 marked 혹스 과정(marked Hawkes process)라고 한다. 위의 식에서는 (2.3)과 비교했을 때 포아송 기본 수준에 해당하는 항이 상수가 아닌 시간의 함수로 표현된다. 뒤에 따라오는 항이 없다면 비동질 포아송 과정의 조건부 강도에 해당한다. 또는 외인성 요인을 배경 자극에 반영할 수도 있다. 배경 자극 $\mu(t)$가 t에 종속적이면, 시간 정상성을 만족하지 않는다 (Gao와 Zhu, 2018). 등호를 기준으로 왼쪽에 있는 항을 살펴보면, 조건부 강도 함수에 비선형 변환을 하여 좀 더 유연한 모형을 만들 수 있는 형태임을 알 수 있다. 비선형 변환을 통해 (2.3)과 달리 $\lambda^*(t)$이 음수 값을 갖게 만들 수 있다. 즉, 앞선 사건이 뒤에 올 사건을 억제하는 효과를 나타낼 수 있다. 사건이 발생하는 위치와 시점을 설명하기 위해 혹스 모형을 시공간 모형으로 확장하여 지진 모형, 범죄 예측, 역학 및 생태학 분석 등에 적용할 수 있다.

3.2. 다변량 혹스 과정

두 가지 이상의 서로 다른 종류의 사건들이 서로 관련이 되어 발생하는 경우가 있다. 미국 증권 시장에서 S&P 500 지수와 다우존스 산업 평균(Dow Jones industrial average) 지수는 비슷한 흐름을 보일 때가 많다. 대표적인 예로는 1987년 10월 19일 뉴욕 증권 시장에서 이 두 지수 모두 대폭락 했던 블랙먼데이 사건이 있었다. 원달러 환율과 유로달러 환율도 서로 관련되어 나타한다. 이렇게 상호적으로 관련이 있는 두 가지 이상의 점 과정을 상호 자극 혹스 과정(mutually exciting processes)으로 분석할 수 있다. 다음과 같은 조건부 강도 함수를 통해 D 가지 종류의 점 과정 $\{N_i(t)\}_{j=1}^D$으로 상호 자극 혹스 과정을 모형화할 수 있다.

$$\lambda_i^*(t) = \mu_i + \sum_{j=1}^{D} \sum_{t_{j:r} < t} g_{ij}\left(t - t_{j:r}\right),$$

여기서, $t_{j:r}$는 j형 r 번째 사건이 일어난 시점이다. $g_{ij}(t - t_{j:r})$은 $t_{j:r}$ 시점에서 j형 사건에 의해 만들어진 i형 사건의 강도에 대한 기여를 나타낸다. 행렬로 표현하면 다음과 같다.

$$\boldsymbol{\lambda}^*(t) = \boldsymbol{\mu} + \int_0^t \boldsymbol{G}(t - \mu)\, d\boldsymbol{N}(u).$$

이 때, 행렬 $\boldsymbol{G}(t-\mu)$는 $g_{ij}(t-u)$를 원소로 한다. 위와 같은 기본 모형은 일변량 자기 자극 모형과 유사한 방식으로 mark, 외인성 요인 및 비선형 변환으로 일반화될 수 있다.

4. 혹스 과정의 적용

4.1. 지진 발생

지진은 여진을 유발하기 때문에 혹스 과정 모형을 이용하여 자기 자극을 반영할 수 있다. Ogata (1988)은 Gutenberg-Richter 법칙과 수정된 Omori 법칙을 기반으로 한 ETAS 모형을 도입했다. ETAS 모형은 지진의 강도를 mark로 한 marked 혹스 과정의 특수한 경우이다. Zhuang 등 (2002)은 다음과 같은 시공간 ETAS 모형을 제안하였다.

$$\begin{aligned} \lambda^*_{\beta,\boldsymbol{\theta}}(t,x,y,m) &= \nu_\beta(m)\,\lambda^*_{\boldsymbol{\theta}}(t,x,y), \\ \lambda^*_{\boldsymbol{\theta}}(t,x,y) &= \mu(x,y) + \sum_{\{i:t_i<t\}} k(m_i)\,g(t-t_i)\,w(x-x_i, y-y_i \mid m_i), \end{aligned} \tag{4.1}$$

여기서 $\nu_\beta(m)$는 mark 함수, $\mu(x,y)$는 배경 자극 함수, $k(m)$은 강도 m인 지진으로 인한 여진의 기대 횟수, $g(t)$는 여진이 발생한 시점의 분포 함수이고 $w(x,y|m)$은 여진의 위치에 대한 분포 함수이다. 배경 자극 $\mu(x,y)$에 대해서는 시간 정상성을 가정하여 지진의 위치(x,y)의 함수로 표현하였다. 각 지진으로부터 독립적으로 자손(offspring)에 해당하는 여진이 발생한다. 앞으로 일어날 여진의 수는 현재 일어난 지진의 강도 m과 관련 있으므로 $k(m)$으로 표시한다. Utsu와 Ogata (1995)에 따르면, 여진의 발생까지의 시간 확률 분포 $g(\cdot)$는 직전에 발생한 지진(direct ancestor)으로부터의 시간 지연 함수(time lag function)이며, 지진의 강도와는 무관하다. 여진의 위치 (x,y)는 직전에 발생한 지진의 강도에 종속적이므로 여진의 위치에 대한 분포 함수 $w(x,y|m)$는 강도 m의 조건부 형태로 나타낸다. 식 (4.1)의 각 항들은 다양한 종류의 식을 이용하여 표현할 수 있다. R-패키지 `ETAS`에는 다음과 같은 식에 대한 추정 방법이 구현되어 있다.

$$\begin{aligned} \nu_\beta(m) &= \beta \exp\{-\beta(m-m_0)\}, \\ \mu(x,y) &= \mu u(x,y), \quad \mu > 0, \\ k(m) &= A\exp\{\alpha(m-m_0)\}, \quad A>0,\ \alpha>0, \\ g(t) &= \frac{p-1}{c}\left(1+\frac{t}{c}\right)^{-p} I_{\{t>0\}}, \quad c>0,\ p>1, \\ w(x,y\mid m) &= \frac{q-1}{\pi\sigma(m)}\left(1+\frac{x^2+y^2}{\sigma(m)}\right)^{-q}, \ \text{with } \sigma(m) = D\exp\{\gamma(m-m_0)\}. \end{aligned} \tag{4.2}$$

ETAS 모형에서 mark 함수 $\nu_\beta(m)$는 주로 지수 분포를 가정한다. $\mu(x,y)$에서 μ는 발생하는 위치와 상관없이 모든 지진이 공통적으로 갖는 상수이고, 지진의 위치와 관련된 항 $u(x,y)$는 smooth function으로 추정한다. 지진 분석에서는 강도가 m_0 이상인 지진만 고려한다. 여진의 기대 횟수 $k(m)$는 지수 함수를 가정하고, 상수 A와 α을 이용하여 유연한(flexible) 형태를 갖도록 하였다. 지진 발생 후 시간이 지날 수록 여진이 발생할 가능성은 낮아지므로, $g(t)$는 Omori의 법칙에 따라 상수 c, p를 이용하여 $(t+c)^p$의 역수에 비례하도록 만든 위와 같은 분포가 주로 이용된다 (Utsu, 1961; Utsu와 Ogata,1995). 여진의 위치에 대한 분포는 상수 D, q, γ를 이용하여 만든 long tail inverse power density이다.

4.2. 범죄 사건의 발생

범죄학 연구에 따르면 범죄는 전염과 같은 과정을 통해 지역 환경을 통해 퍼질 수 있다. 예를 들어, 어떤 지역의 보안이 취약한 것을 잘 아는 강도들이 비슷한 수법으로 인근 지역을 반복적으로 강탈하는 경향이 있다. 또한

갱(gang) 집단 간에 서로 보복하는 폭력 사건이 빈번하게 발생한다. 따라서, 범죄 사건은 한 범죄 사건이 다른 범죄 사건을 유도하는 것으로 볼 수 있으므로, 자기 자극을 포함한 조건부 강도 함수를 이용한 혹스 과정으로 설명할 수 있다. Egesdal 등 (2010)은 미국 로스엔젤레스 지역에서 발생하는 갱 집단간의 폭행 사건의 패턴을 혹스 과정을 통해 모형화하였다. Egesdal 등 (2010)는 시간만 고려한 혹스 과정을 이용하였으며, 지수붕괴 함수 (3.1)로 자극 함수를 표현하였다. Mohler 등 (2011)는 다음과 같이 위치를 표시하는 변수 x, y를 추가하여 시공간 상에서 자기 자극을 반영한 혹스 과정을 구현하였다.

$$\lambda^*(t,x,y) = \nu(t)\,\mu(x,y) + \sum_{\{k:t_k<t\}} g(t-t_k,\, x-x_k,\, y-y_k).$$

Mohler 등 (2011)는 자극 함수 g에 대해서 비모수 모형으로 분석하는 것을 권장하였으며, 비모수 추정 방법을 이용하여 로스앤젤레스 경찰서에서 제공한 주거 절도 데이터를 분석하였다.

4.3. 전염병 모형

전염성 질병은 그 질병에 감염된 사람으로부터 전파되기도 하고, 또는 개인이 인수공통 전염병 확산과 같은 외부 요인으로부터 독립적으로 감염되거나 이미 감염된 지역 사회로 여행함으로써 감염될 수도 있다. 이런 점에서 혹스 과정은 전염병 모형에 유용하게 이용될 수 있다. 전염병 모형에서 자주 쓰이는 모형으로는 susceptible-infected-recovered (SIR) 모형이 있는데, Rizoiu 등 (2018)은 SIR 모형과 혹스 모형을 혼합하여, 전염병 감염의 비율 및 회복률을 반영한 SIR 혹스 모형을 제시하였다. 혹스 모형에서는 발생하는 사건의 수에 대한 상한선이 없지만, 전염병 확산을 비롯하여 많은 경우에 사건은 유한하게 일어난다. Rizoiu 등 (2018)가 제시한 모형은 사건 횟수의 유한성을 고려하여 만든 모형으로 HawkesN 모형이라고 부른다.

조건부 강도 함수는 다음과 같다.

$$\lambda^H(t) = \left(1 - \frac{N(t)}{N}\right)\left\{\mu + \sum_{t_i<t} g(t-t_i)\right\}. \tag{4.3}$$

이 때, $g(t-t_i)$는 혹스 과정에서 정의한 자극 함수이다. N은 총 인구수인데 상수로 가정하고, $N(t)$는 감염자에 대한 계수 과정이다. $\lambda^H(t)$와 $N(t)$는 right-continuous인 함수이다. $1-N(t)/N$ 항은 t 시점 이후에 여전히 발생할 수 있는 사건의 비율이고, t 시점에서의 사건 발생 비율을 조정한다. $t=0$일 때에는 $\lambda^H(t)=\mu$이고 $N(t)=N$이면, $\lambda^H(t)=0$이 되어 더 이상 감염이 발생하지 않는다. N이 무한히 큰 값을 가질 때, SIR 혹스 과정은 혹스 과정 (2.3)과 같게 된다. Rizoiu 등 (2018)는 다음과 같은 자극 함수를 이용하였다.

$$g(t) = \kappa\theta e^{-\theta t}.$$

이 때, κ는 scaling factor이고, θ는 지수 함수의 모수이다.

Unwin 등 (2021)은 SIR 혹스 모형 (4.3)에서 배경 자극 μ를 t에 대한 함수 $\mu(t)$로 변경하고, 자극 함수에 대해서는 Rayleigh kernel를 이용하여 말라리아 전염에 대하여 연구하였다.

$$g(t-t_i) = \alpha(t-t_i)\,e^{-\frac{\delta(t-t_i)^2}{2}}, \quad t>t_i,$$

여기서 $\alpha>0$은 감염된 개인의 감염력의 정도, $\delta>0$는 감염 기간을 조절하는 모수이다. $\mu(t)$에 대해서는 말라리아 발병에 대한 계절성을 반영하기 위하여 consine과 sine을 이용하여 cycle을 갖는 함수를 이용하였다.

4.4. 금융

금융 분야에서는 선물 시장에서의 거래 건 수와 거래가 이루어지는 시점 및 주가의 변동성 등을 모형화하기 위해 혹스 과정이 이용될 수 있다. 거래 수준의 변동성은 거래, 중간 가격 변동 등의 사건의 발생의 수와 직접적으로 관련될 수 있으므로 혹스 과정의 자기 자극 특성으로 이러한 변동성을 설명하고, 변동성의 클러스터링을 통해 자기 자극으로 인한 변동이 시작된 시점을 추정할 수 있다. Bowsher (2007)는 단변량 혹스 과정을 보정하여 좀 더 일반화된 다변량 사건 데이터에 대한 연속 시간을 고려한 혹스 모형을 제시하고, NASDAQ 및 NYSE의 일중 주식 데이터를 분석하였다. Bacry 등 (2012)는 2009년 75 거래일 동안 발생한 10년 Euro-Bund 선물 계약에 대해 보정된 단변량 혹스 과정을 이용하였고, 비모수 방법으로 자극 함수를 추정하였다. Abergel과 Jedidi (2015)는 지수 커널을 사용하는 다변량 혹스 과정으로 호가창(limit order book; LOB)에서의 다양한 거래, 취소 등의 도착에 대한 모형을 제시하였다. Bacry와 Muzy (2016)은 EUREX 거래소에서 두 개의 유동적인 선물 자산의 시장 주문에 대해서 멱법칙 자극 함수를 이용한 혹스 과정으로 설명하였다. Rambaldi 등 (2017)에서는 다변량 혹스 과정 모형에 비모수 추정 방법을 도입한 방법으로 S&P 500에서 LOB의 주문 도착 시간과 주문 크기 사이의 복잡한 상호 작용을 분석하였다.

5. 혹스 과정의 추정

4.1절에서 설명한 ETAS 모형 (4.1)를 예를 들어 설명하고자 한다. 로그 가능도 함수는 다음과 같다.

$$l(\boldsymbol{\theta}) = \sum_i \lambda^*(t_i, x_i, y_i) - \int \lambda^*(t, x, y)\, dx\, dy\, dt.$$

위의 식에서 로그 가능도 함수가 명시적으로 표현되지만 numerical 방법을 이용한 최대 가능 추정치(maximum likelihood estimator; MLE)의 계산은 간단하지 않다. 첫 번째 이유는, 발생하는 사건은 셀 수 있지만, 그 사건이 앞선 사건에 의해서 자극이 되어 일어나는 것인지 독립적으로 일어나는 것인지 알 수 없으므로 배경 자극의 정확한 값을 알 수 없는 incomplete data이다. 예를 들어, 지진 데이터의 경우, 본진과 여진을 데이터만으로는 구별할 수 없다. 두 번째로, 로그 가능도 함수의 최대값을 포함한 영역이 극도로 flat한 경우가 많아서 MLE를 구하는 데 어려움이 생길 수 있다. Zhuang 등 (2002)에서는 ETAS 모형 (4.1)에 대하여 (4.2)의 배경 자극 $\mu(x, y)$ 을 반복적으로 업데이트하여 MLE를 찾는 방법을 제시하였으나, 이 방법은 초기값에 따라 MLE가 달라질 수 있다.

Veen과 Schoenberg (2008)는 numerical 방법의 문제점으로 초기값에 따라 MLE가 달라질 수 있음을 지적하고, EM 알고리즘 방법을 통해 추정치를 구하는 방법을 제시하였다. ETAS 모형에서 (4.2)의 배경 자극과 달리 $\mu(x, y)$을 포아송 분포를 이용하여 추정한 후, 이 추정치를 이용하면 complete data에 대한 로그 가능도 함수 $l_c(\boldsymbol{\theta})$를 표현하였다. i 번째 지진이 앞서서 일어난 j 번째 지진에 의해 발생된 지진일 때, 이것을 $u_i = j$로 표현하기로 하자. E-step에서는 확률 $P(u_i = j)$를 추정하고, M-step에서는 $P(u_i = j)$의 추정치를 이용하여 $l_c(\boldsymbol{\theta})$를 최대화하는 과정을 반복하여 $\boldsymbol{\theta}$의 추정치를 구할 수 있다.

Rasmussen (2013)은 혹스 과정의 추정 방법으로 베이지안 방법을 제시하였다. Rasmussen (2013)은 배경 자극(μ 또는 $\mu(t)$)과 자기 자극에 의해 발생한 사건의 수가 주어져 있지 않기 때문에, Metropolis-within-Gibbs 알고리즘을 이용하여 $\mu(t)$를 추정하고, 이 값을 이용하여 모수에 대한 사후 분포를 구하는 방법을 도입하였다. 이 과정을 반복하여 가능도 함수를 최대화하는 추정치를 찾을 수 있다. ETAS 모형 분석시 Omi 등 (2015)에서는 Metropolis-within-Gibbs 알고리즘 대신에 지진학에서 주로 쓰이는 Gutenberg-Richter 공식 또는 앙상블 러닝을 이용하여 본진과 여진에 대한 예측을 한 후, 모수의 사후 분포를 구하였다. 각 모수의 사전 분포 선택에 대한 설명은 Ross (2021)를 참고하기 바란다.

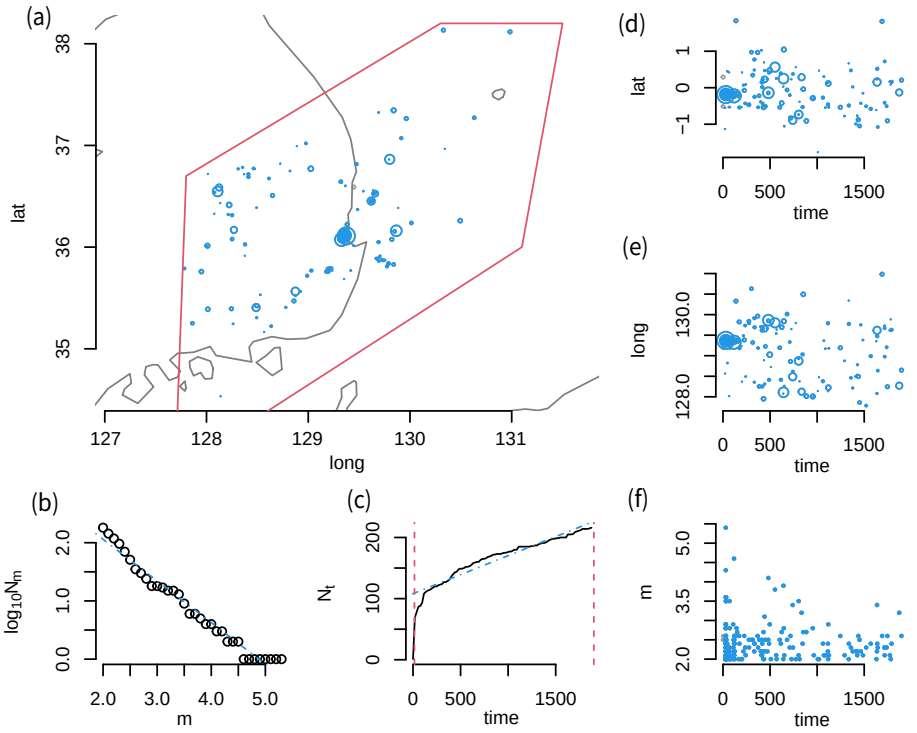
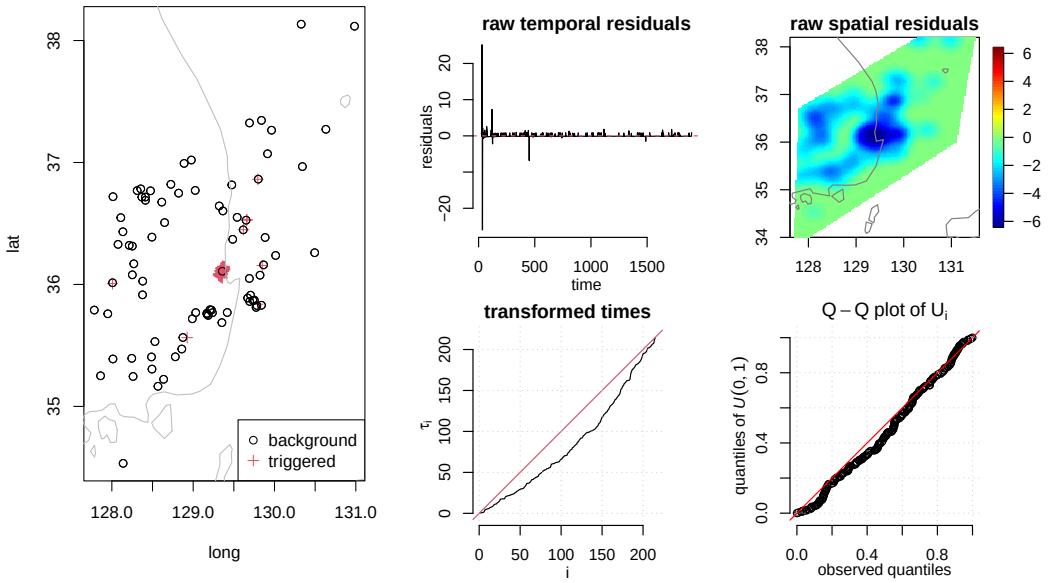


Figure 1: *Results of ETAS for Gyeongsang provinces in Korea. (a) location of epicenters, (b) logarithm of frequency by magnitude, (c) cumulative frequency over time, (d) latitude against time, (e) longitude against time, (f) magnitude against time.*

6. 경상도 지진 분석

기상청 기상자료개방포털에 공개된 데이터를 이용하여 2017년 11월부터 2022년 12월까지 경상도에서 발생한 지진을 혹스 과정 모형으로 분석하였다. 모형 (4.1)의 (4.2) 식을 가정하고, R-패키지 `ETAS`를 사용하였다. 참고로, 일반적인 혹스 과정과 관련된 R-패키지로는 `hawkes`와 `hawkesbow`가 있다. `hawkes` 패키지는 혹스 과정의 시뮬레이션은 가능하나 모수 추정 함수는 제공하지 않고, `hawkesbow` 패키지는 시뮬레이션 및 모수 추정 함수를 제공하지만, 자극 함수가 지수붕괴 함수와 멱법칙 함수인 경우에 국한되므로 ETAS 모형의 추정에는 사용할 수 없다.

Figure 1은 `catalog` 함수에 의해 생성되었다. Figure 1(a)는 경상도에서 일어난 진원을 동그라미로 표시하였고, 동그라미의 크기는 지진의 강도에 비례한다. 2017년 11월 15일 포항시에서 일어난 규모 5.4의 지진은 지도에서 가장 큰 동그라미로 표시되어 있다. Figure 1(b)에는 지진의 강도(m)에 따른 $\log_{10} N_m$을 표시하였다. N_m은 지진의 강도 m보다 크거나 같은 크기의 지진의 수를 뜻한다. Gutenberg-Richter 법칙은 지진의 강도와 지진의 횟수의 관계를 설명하는 법칙으로, $\log_{10} N_m = a - bm$ (a, b는 상수)의 식으로 표현된다 (Ogata, 1998). m 대비 $\log_{10} N_m$의 그래프가 선형으로 적합이 되면 데이터가 Gutenberg-Richter 법칙을 잘 따르는 것으로 볼 수 있다. Figure 1(b)에서는 강도 5이상의 지진을 제외하면 전반적으로 직선에 가까운 패턴을 보이므로 이 데이터는 해당 지역의 지진을 분석하는 데 적합하다고 볼 수 있다. Figure 1(c)에는 시간에 따른 누적된 지진의



(a) Background and triggered events (b) Diagnostic plots for the fitted ETAS model

Figure 2: *The results of ETAS model. (a) background and triggered events detected by estimated declustering probabilities. (b) diagnostic plots for the fitted ETAS model: the temporal residuals (top left), smoothed spatial residuals (top right), transformed times τ_i against i (bottom left) and the uniform Q-Q plot of U_i (bottom right).*

수가 표시되어 있으며, 지진 발생의 시간 정상성을 파악하는 데 이용된다. 시간 정상성 조건 하에서는 mark와 사건이 일어나는 시점은 독립이고, 시간 간격 T에서 기대되는 사건의 수는 T에 비례한다. 이 그래프가 직선에 가까우면 시간 정상성을 따른다고 볼 수 있다. Figure 1(c)은 직선에서 크게 벗어나지 않았으므로 시간 정상성을 따른다고 볼 수 있다. Figure 1(d)–(f)에는 각각 지진이 발생한 시간에 따른 지진의 위도, 경도, 지진의 강도를 표시하였다. Figure 1(f)에서는 강도가 큰 지진 후에 우하향하는 방향으로 점들이 연이어 찍혀 있는 것은 본진 이후 여진이 발생했을 가능성이 있다는 것을 의미한다.

ETAS 모형 (4.1)에서 추정해야하는 모수는 β와 $\boldsymbol{\theta} = (\mu, A, c, \alpha, p, D, q, \gamma)$이다. R에서 `etas` 함수를 사용하여 다음과 같은 결과를 얻었다.

```
ETAS model: fitted using iterative stochastic declustering method
converged after 5 iterations: elapsed execution time 1.02 minutes

ML estimates of model parameters:
           beta     mu      A      c  alpha      p      D      q  gamma
Estimate 2.4239 1.0577 0.1932 0.0140 1.4892 1.1671 0.0001 2.1894 0.6107
StdErr   0.0273 0.0264 0.0648 0.1198 0.0235 0.0103 0.1743 0.0406 0.0885

Declustering probabilities:
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0304  0.4522  0.9987  1.0000

log-likelihood:  142.3977  AIC:  -268.7954
```

$p > 1, \beta > \alpha$이고 $\rho = A\beta/(\beta - \alpha) < 1$은 정상성 조건에 대한 충분조건이다 (Daley와 Vere-Jones, 2003, p204). 이 데이터는 ρ의 추정치가 대략 0.50으로, 시간 정상성 조건을 만족한다.

발생한 지진을 추정된 모형을 이용하여 본진과 여진으로 구분할 수 있다. 위의 추정 결과에서 계산된 Declustering probabilities 값에 근거하여, Figure 2(a)에 본진과 여진일 가능성이 95%이상인 지진에 대하여 각각 동그라미와 십자가로 표시하였다. Figure 2(b)는 `resid.etas` 함수를 이용하여 그린 그림으로 잔차에 대한 diagnostic plot을 보여준다. 11월 15일(time = 15)에 포항시에서 일어난 규모 5.4의 지진은 다른 지진과는 패턴이 다른 특별한 사건으로 보여진다. 왼쪽 상단 잔차 그림에서도 이 지진(time = 15)을 제외하면 전반적으로 특별한 패턴은 보이지 않는다. 오른쪽 상단 공간 잔차 그림에서도 포항시 부분만 큰 값을 갖는 것을 알 수 있다. Ogata (1988b)에서는 다음과 같이 조건부 강도 함수의 적분값을 transformed time τ로 정의하고, 그 패턴을 확인하는 분석 방법을 제시하였다.

$$\tau_i = \int_0^{t_i} \lambda^*(t)\, dt.$$

추정된 조건부 강도 함수의 적분 값은 누적된 사건 발생 횟수에 대한 추정값이다. 추정된 강도 함수를 이용하여 구한 transformed time과 실제로 누적된 발생횟수가 비슷한 값이면, 추정된 강도 함수가 실제 데이터를 잘 반영한다고 볼 수 있다. Figure 2(b) 왼쪽 하단에 제시된 사건 대비 transformed times 그래프는 직선을 따르지 않고 직선 아랫쪽에 곡선 형태를 보이고 있으며, 전반적으로 실제 누적 발생 건수보다 추정치가 더 작은 값을 갖는다. 즉, 모형으로 설명되지 않은 noise가 존재한다. Figure 2(b) 오른쪽 하단에는 Q-Q plot이 제시되어 있다. 여기서 U_i는 다음과 같이 계산된 값이다.

$$U_i = 1 - \exp\{-(t_i - t_{i-1})\}.$$

이 값은 연이은 지진의 발생 시간 간격을 지수함수를 이용하여 변환한 값인데, 이 값의 quantile은 시간 정상성을 따를 경우 Uniform (0,1) 분포의 quantile과 직선 관계에 있다. U_i의 Q-Q plot은 직선에 가까운 패턴을 보이고 있으므로, 시간 정상성을 갖는다. Kolmogorov-Smirnov 검정을 통해 Uniform (0,1) 분포와 동일 여부를 검정해보았을 때에도 p-value가 0.6362로 Uniform (0,1) 분포와 다르다고 할 수 없으므로 시간 정상성을 갖는다고 볼 수 있다.

7. 결론

1971년에 혹스 과정이 소개된 이후, 초반에는 주로 지진 발생에 적합한 ETAS 모형이 소개되고 발전되었으나, 최근 몇 년에 걸쳐서 혹스 과정의 새로운 이론적 발전과 다양한 분야의 적용이 급속도로 진행되었다. 혹스 과정을 구성하는 배경 자극과 자극 함수를 사건 발생의 패턴에 맞게 선택하여 지진 뿐만 아니라, SNS를 통한 소식 및 영상 확산, 전염병 모형화 및 범죄 발생 패턴 분석 등 다양한 분야에 활용 가능하다. 모수가 많은 복잡한 모형일수록 모수의 추정치를 구하기 어려운 점이 있다. R-패키지 `ETAS`에는 몇몇 ETAS 모형에 대해서 numerical 방법으로 구현된 최대가능 추정 함수를 제공하고 있으나, 추정량이 초기값에 민감하다는 단점이 있다. EM 알고리즘을 이용한 추정 방법이 numerical 방법보다 좀 더 안정적인 것으로 알려져 있고, 베이지안 방법도 적용 가능하다.

References

Abergel F and Jedidi A (2015). Long-time behavior of a hawkes process–based limit order book, *SIAM Journal on Financial Mathematics*, **6**, 1026–1043.

Bacry E, Dayri K, and Muzy JF (2012). Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data, *The European Physical Journal B*, **85**, 1–12.

Bacry E and Muzy JF (2016). First-and second-order statistics characterization of Hawkes processes and non-parametric estimation, *IEEE Transactions on Information Theory*, **62**, 2184–2202.

Bowsher CG (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models, *Journal of Econometrics*, **141**, 876–912.

Daley DJ and Vere-Jones D (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*, Springer, New York.

Daley DJ and Vere-Jones D (2008). *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*, Springer, New York.

Egesdal M, Fathauer C, Louie K, Neuman J, Mohler G, and Lewis E (2010). Statistical and stochastic modeling of gang rivalries in Los Angeles, *SIAM Undergraduate Research Online*, **3**, 72–94.

Gao X and Zhu L (2018). Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues, *Queueing Systems*, **90**, 161–206.

Hawkes AG (1971). Spectra of some self-exciting and mutually exciting point processes, *Biometrika*, **58**, 83–90.

Hawkes AG (2018). Hawkes processes and their applications to finance: A review, *Quantitative Finance*, **18**, 193–198.

Hawkes AG and Oakes D (1974). A cluster process representation of a self-exciting process, *Journal of Applied Probability*, **11**, 493–503.

Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, and Tita GE (2011). Self-exciting point process modeling of crime, *Journal of the American Statistical Association*, **106**, 100–108.

Ogata Y (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *Journal of the American Statistical Association*, **83**, 9–27.

Ogata Y (1998). Space-time point-process models for earthquake occurrences, *Annals of the Institute of Statistical Mathematics*, **50**, 379–402.

Ogata Y and Katsura K (1988). Likelihood analysis of spatial inhomogeneity for marked point patterns, *Annals of the Institute of Statistical Mathematics*, **40**, 29–39.

Omi T, Ogata Y, Hirata Y, and Aihara K (2015). Intermediate-term forecasting of aftershocks from an early aftershock sequence: Bayesian and ensemble forecasting approaches, *Journal of Geophysical Research: Solid Earth*, **120**, 2561–2578.

Rambaldi M, Bacry E, and Lillo F (2017). The role of volume in order book dynamics: A multivariate Hawkes process analysis, *Quantitative Finance*, **17**, 999–1020.

Rasmussen JG (2013). Bayesian inference for Hawkes processes, *Methodology and Computing in Applied Probability*, **15**, 623–642.

Rizoiu MA, Mishra S, Kong Q, Carman M, and Xie L (2018). SIR-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations, In *Proceedings of the 2018 world wide web conference*, Lyon, France, 419–428.

Ross GJ (2021). Bayesian estimation of the ETAS model for earthquake occurrences, *Bulletin of the Seismologi-*

cal Society of America, **111**, 1473–1480.

Unwin HJT, Routledge I, Flaxman S *et al.* (2021). Using Hawkes processes to model imported and local malaria cases in near-elimination settings, *PLoS Computational Biology*, **17**, e1008830.

Utsu T (1961). A statistical study of the occurrence of aftershocks, *Geophysical Magazine*, **30**, 521–605.

Utsu T and Ogata Y (1995). The centenary of the Omori formula for a decay law of aftershock activity, *Journal of Physics of the Earth*, **43**, 1–33.

Veen A and Schoenberg FP (2008). Estimation of space–time branching process models in seismology using an em–type algorithm, *Journal of the American Statistical Association*, **103**, 614–624.

Zhuang J, Ogata Y, and Vere-Jones D (2002). Stochastic declustering of space-time earthquake occurrences, *Journal of the American Statistical Association*, **97**, 369–380.

Received February 1, 2023; Revised March 12, 2023; Accepted March 13, 2023

혹스 과정의 개요 및 응용

김미정[1,a]

[a]이화여자대학교 통계학과

요 약

혹스 과정은 자기 자극 특성을 가진 점 과정으로서, 지진 발생시 본진으로 인한 여진이 발생되는 현상을 설명하는 데 주로 쓰이는 확률 모형이다. 최근에는 전염병 확산, SNS에서의 소식 확산 등 자기 자극을 특성을 가진 다양한 현상을 설명하는 데 활용되고 있다. 혹스 과정은 다양한 형태의 자극 함수를 도입하여 발생하는 사건의 특성에 따라 유연하게 변형이 가능한데, 최대 우도 추정량을 구하는 것이 쉽지 않기 때문에 최근까지도 개선된 추정 방법이 제시되고 있다. 이 논문에서는 혹스 과정을 설명하기 위해 조건부 강도 함수와 자극 함수에 대해 설명하고, 지진, 전염병, 범죄 및 금융에서 활용되었던 예와 추정 방법을 알아보도록 한다. R-패키지 ETAS를 이용하여 2017년 11월부터 2022년 12월까지 한국 경상도에서 발생한 지진을 분석하도록 한다.

주요용어: 자기 자극, 자극 함수, 조건부 강도 함수, 지진 모형, 혹스 과정

이 논문은 연구재단 연구 과제 NRF-2020R1F1A1A01074157에 의하여 수행되었음.

[1](03760) 서울시 서대문구 이화여대길 52, 이화여자대학교 통계학과. E-mail: m.kim@ewha.ac.kr