Journal of the Korean Data & Information Science Society
2015, **26**(2), 301–312

http://dx.doi.org/10.7465/jkdi.2015.26.2.301
한국데이터정보과학회지

확률적 방법에 기반한 질병 확산 모형의 구축[†]

류수락[1] · 최보승[2]

[1]대구대학교 대학원 통계학과 · [2]대구대학교 전산통계학과

접수 2015년 1월 5일, 수정 2015년 1월 29일, 게재확정 2015년 2월 10일

요 약

본 연구는 전염병의 확산 과정을 설명하기 위한 질병 확산 모형을 구축 하고자 하였다. 질병의 확산 과정은 결정적인 과정과 확률적인 과정으로 크게 분류할 수 있다. 대부분의 연구가 질병의 확산 과정을 결정적 과정으로 움직인다고 가정을 하고 상미분방정식을 이용하여 모형을 구축하였다. 본 연구에서는 질병 확산 모형인 SIR (Suspectible – Infectious – Recovered) 모형을 기반으로 하여 질병의 확산 예측 모형을 구현하고자 하였다. 최소제곱법을 이용하여 모수를 추정한 후, 상미분방정식을 이용한 결정적 모형 방법과 더불어 Gillespie가 제안한 방법에 기반하여 확률적인 과정을 따르는 모형 적합을 함께 시도하였다. 본 연구에서 소개된 방법들은 질병관리본부의 2001년 1월부터 2002년 3월까지의 국내 말라리아 주별 발병자 수 자료를 이용하여 모형 적합을 시도 하였으며, 그 결과 구현된 모형이 실제 질병의 확산과정을 잘 설명하였다.

주요용어: 길레스피 알고리즘, 상미분방정식, 에스아이알 모형, 질병 확산 모형, 확률적 반응 모형.

1. 서론

전염병이란 질병 중 전염이 가능한 질병을 말한다. 특정 병원체나 병원체의 독성물질로 인하여 발생하는 질병으로 감염된 사람으로부터 감수성이 있는 숙주 (사람)에게 감염되는 질환을 의미한다. 전염병 병원체의 종류로는 세균, 바이러스, 기생충, 곰팡이, 원생동물 등이 있으며 임상 특성으로는 호흡기계 질환, 위장관 질환, 간질환, 급성 열성 질환 등이 있다. 확산 방법은 사람간 접촉, 식품이나 식수, 곤충 매개, 동물에서 사람으로 확산, 성적 접촉 등에 의한다. 실제로 겨울철에 사람들이 많이 걸리는 감기의 주된 원인은 바이러스다. 주로 호흡기를 통해 전염되며 손이나 신체 일부의 접촉에 의해서도 옮는다. 접촉성 바이러스는 대개 사람과 사람이 만나야 전염된다.

2011년 미국에서 개봉된 영화 <컨테이젼>과 2013년 한국에서 개봉된 영화 <감기>는 전염병의 공포를 그리고 있다. <컨테이젼> 영화 속 공포의 원인은 박쥐의 병균이 돼지로 옮아 생긴 치명적인 바이러스였다. 이 바이러스는 상상을 초월한 속도로 퍼져나갔고, 많은 사람들이 이 바이러스에 전염되어 손을 쓸 사이도 없이 죽었다. 영화 <감기>는 호흡기로 감염되며 감염속도 초당 3.4명, 치사율 100%의 유례 없는 최악의 바이러스가 대한민국에 발병하였다. 이에 정부는 전세계적인 확산을 막기 위해 국가재난사태를 발령하였다. 이처럼 전염병의 공포는 영화에만 있지 않다. 신종플루, 사스 (SARS) 등 많은 사람들이 감염되었고, 구제역으로 수많은 동물이 죽었다.

[†] 이 논문은 2014년도 정부 (교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (No.2012R1A1A1010156).
[1] (712-714) 경상북도 경산시 진량읍 대구대로 201, 대구대학교 대학원 통계학과, 석사과정.
[2] 교신저자: (712-714) 경상북도 경산시 진량읍 대구대로 201, 대구대학교 전산통계학과, 조교수.
E-mail: bchoi@daegu.ac.kr

실제 1918년에 유행했던 스페인독감으로 인해 2,500만~5,000만명이 희생되었다. 미국의 한 신병 훈련소에서 처음 발생한 뒤 1차 세계대전 중 세계 각지로 이동한 군인들에 의해 18개월 만에 세계 인구의 20%가 감염되었고, 그 중 2.5%~5%가 사망했다. 이때는 세계적인 항공망이 제대로 갖춰지기 전이라 전파 속도가 상대적으로 느렸다. 하지만 이젠 누구나 비행기를 타고 세계 어느 곳으로든 하루 안에 이동할 수 있어 전염병이 아주 빠르게 퍼진다. 또한 도시에 사람이 엄청나게 밀집해 있어 전염병이 어떻게 퍼질지 알기 힘들다. 전염병을 막기 위해 전 세계의 대부분 국가들은 매우 오랫동안 전염병의 발생과 확산을 예측하고 대비하기 위해 전염병 모형을 사용하고 있다. 이런 연구 성과를 이용하여 과학자들은 전염병 발생시 예방접종과 격리에 대한 계획을 세우거나 특별한 전염병에 대한 치사율을 추정하는데 적용하려고 노력하였다.

전염병의 전이 또는 확산 과정을 모형화 함으로써 전염병이 발생하였을 때 이를 통제하기 위한 초기 대응방안으로 이용할 수 있으며 또한 시간의 흐름에 따른 병의 확산 과정을 예측할 수 있다. 이와 같은 전염병의 확산 과정을 모형화 함으로써 사망의 원인을 정량화 하려는 최초의 연구는 Graunt (1662)에 의해 진행되었다. 이 연구에서는 매주 마다 사망자들의 원인과 수를 목록화하고 연구하였다. Bernoulli (1766)는 천연두를 연구하여 이 전염병 때문에 사람들이 얼마나 죽었는지 분석한 결과를 발표했으며, 이 연구를 통하여 천연두 균으로 인해 인체에 면역력을 갖게 하므로 천연두를 예방할 수 있다는 주장을 뒷받침하게 되었다. 그 이후 전염병 모형을 체계적으로 연구하기 시작한 것은 20세기 초부터이다. Hamer (1906)는 영국 런던에서 발생했던 홍역의 유행에 관한 모형인 SI 모형을 제시하였고, Ross (1911)는 말라리아는 모기가 옮기는 병이라는 것을 알아내고 말라리아의 확산 모형을 제시함으로써 말라리아 예방에 기여하였다.

전염병의 확산 과정을 확률적으로 모형화 하기 위한 방법으로 Kermack와 McKendrick (1927)은 전염병이 유행하기 위한 초기 조건과 전염병의 확산 정도를 예측하기 위하여 제안한 SIR 모형이 있다. SIR 모형은 전염병의 전이 상태를 크게 세 가지로 구분하였다. 전체 모집단 가운데 질병에 감염된 가능성이 있는 집단을 S (susceptible)로 나타낸다. 질병에 감염된 집단은 I (infected)로 나타낸다. 마지막으로 질병으로 회복되거나 사망 등으로 질병으로부터 벗어난 집단은 R (recovered)로 나타낸다. SIR 모형에서는 S 상태에 놓여있는 사람 (개체)이 I 상태와 R 상태로 차례차례로 전이되어 간다는 가정에서 질병의 확산과정을 수학적으로 설명하고자 한 모형이라 할 수 있다. 이 연구는 20세기 중반에 들어와서 폭발적으로 성장하였다. 이후 다양한 수학적인 모형이 만들어졌고, 분석되었으며 실제 전염성 질병에 응용되었다.

전염병 모형을 주제로 하는 연구가 현재까지 우리나라에서도 많이 진행되어 왔다. Hwang 등 (2007)에서는 한국의 말라리아, 신증후군 출혈열, 홍역 자료에 비선형 회귀식으로 표현되는 SIR 모형을 적용하여 기존 현상을 설명하고 미래를 예측하는 연구를 하였고, Lee 등 (2009)에서는 후향연산식을 이용하여 국내 쯔쯔가무시증의 감염자 분포 추정과 질병 확산 모형인 SIRS 모형을 적용하여 유병자수를 추정 하였다. SIRS 모형은 SIR 모형과 달리 recovered 단계에서 다시 일부는 susceptible 단계로 이동이 가능함을 가정하는 모형이다. Lee 등 (2010)은 신종 인플루엔자에 대하여 SIR 기본 모형에서 잠복기 (exposed)를 고려하고 전체 모집단이 닫혀 (closed)있지 않고, 인구의 생성 (birth)과 사망 (death)과정을 모형에 추가하는 SEIR-BD 모형을 이용하여 모형화 하였다. Kim 등 (2013)은 수학적 모형인 SIR 모형을 확장하여 신종인플루엔자 A의 실시간 감시 및 관리를 위해 전염병의 감염 경로 추적 및 예측할 수 있는 통합 정보 시스템을 제안하였다. Hwang과 Oh (2014)는 2010/2011년도 한국 발생 구제역 확산에 관한 연구로 2010/2011 구제역에 대하여 시간-공간 확률 SIR 확률모형을 가정하고 시간과 공간에 따르는 전파 현상에 대하여 고찰하였다. 하지만 대부분의 연구가 전염병의 확산 과정이 결정적 (deterministic)으로 움직인다고 가정을 하고 상미분방정식 (ordinary differential equation; ODE)을 이용하여 모형을 구축하였다. 전염병의 확산 과정은 결정적인 과정으로 움직인다고 볼 수 있지

만, 최근 연구에서는 전염병의 확산 과정을 확률적 (stochastic)인 과정에 따라 확산한다고 보고 모형 구축 연구가 진행되고 있다 (Choi와 Rempala, 2012). Choi와 Rempala (2012)는 2009년에 발생한 신종 인플루엔자 (H1N1)의 확산 과정을 확률적 방법으로 모형을 적합하였다.

이에 본 연구에서는 전염병 모형 중 SIR 모형을 사용하여 전염병의 확산 예측 모형을 구현하고자 하였고, 질병의 확산 과정을 결정적 (deterministic)인 과정과 함께 확률적 (stochastic)인 과정에 따라 확산 한다고 보고 상미분방정식을 이용한 결정적 모형 방법과 더불어 Gillespie (1977)가 제안한 방법에 기반하여 확률적 과정을 따르는 모형 구축을 함께 시도하였다.

본 연구의 진행은 다음과 같다. 먼저 2절에서는 SIR 모형에서 질병의 확산 과정을 결정적 (deterministic)인 과정에 따라 확산한다고 보고 상미분방정식을 이용한 고전적 질병 확산 모형을 소개한다. 그리고 3절에서는 질병의 확산 과정을 확률적 (stochastic)인 과정에 따라 확산한다고 보고 확률적 질병 확산 모형과 Gillespie (1977)에 의해 고안된 확률 시뮬레이션 알고리즘 (stochastic simulation algorithm)을 소개한다. 4절에서는 실제 자료를 이용한 분석결과를 제시하여 제안된 방법이 잘 적용됨을 실험을 통해 보일 것이다. 마지막 5절에서는 결론으로 본 연구의 방법의 한계점과 추후 진행방향에 대하여 논하고자 한다.

2. 고전적 질병 확산 모형

전염병은 인구 집단 내에서 일정한 확산 속도로 퍼져 나가며 전염병에 영향을 끼치는 요소들을 수학적인 전염병 모형으로 만들 수 있다. 수학적인 전염병 모형으로는 병리학적인 특성에 따라 SIR, SEIR, SIRS 등이 있으며 이러한 모형들은 질병의 진행과정을 나타낸다. 대표적인 전염병 모형으로는 Kermack와 McKendrick (1927)이 제시한 모형으로 전염병이 유행하기 위한 초기 조건과 전염병의 확산 정도를 수학적으로 묘사한 SIR 모형이다.

SIR 모형은 질병 확산 모형에서 많이 사용되는 모형으로 전체 모집단을 질병에 걸릴 가능성이 있는 감염 가능군 (suspectible; S), 이미 질병에 감염된 감염군 (infected; I) 그리고 일정 시간이 흐른 후에 질병으로부터 회복되거나 (recovered) 또는 사망 (removed)된 회복군 (recovered; R)으로 구분하고, 각 개체들은 전체 모집단에서 Suspectible → Infectious → Recovered (SIR)의 순으로 전이되어 간다고 가정한다. 회복군으로 이동한 개체 혹은 사람은 면역이 생겨서 더 이상 질병에 감염되지도 않고 다른 사람에게 전염을 시키지도 않는다고 가정한다.

이와는 다르게 SIRS 모형은 회복군에서 일부가 다시 감염 가능군으로 전이될 수 있음을 가정하게 되며 Suspectible → Infectious → Recovered → Suspectible (SIRS)의 순으로 이동한다고 할 수 있다. SIRS 모형에서는 질병으로 회복되어 면역이 생기더라도 다시 질병에 감염될 수 있음을 의미한다. SEIR 모형은 감염 가능군에서 감염군으로 전이되는 과정에서 질병의 잠복기 (exposed)를 추가적으로 고려하는 모형이다. 즉, Suspectible → Exposed → Infectious → Recovered (SEIR)의 순으로 이동한다고 할 수 있다. 본 연구에서는 가장 고전적인 형태인 SIR 모형을 중심으로 연구를 진행하고자 한다.

이제 관찰된 시점 t에서 각 군의 상태를 각각 $X(t)$, $Y(t)$, $Z(t)$로 표현한다. 고전적인 SIR 모형에서는 전체 모집단의 크기는 고정되어 있다고 가정한다. 전체 모집단의 크기를 M이라 하면 모든 시점 t에서 $X(t) + Y(t) + Z(t) = M$이 성립한다.

마지막으로 SIR 모형을 구현하기 위해서는 두 개의 모수를 정의한다. 첫 번째 모수는 질병의 감염 정도를 나타내는 감염률 (transmission rate)이고 또 다른 하나는 회복 정도를 나타내는 회복률 (recovery rate)이다. 이를 각각 λ와 γ로 표현한다. 궁극적으로 SIR 모형을 구축한다고 함은 주어진 데이터로부터 이 두 모수를 추정함으로써 구현된다고 할 수 있다.

SIR 모형에서 질병의 확산 과정을 결정적 (deterministic)으로 움직인다고 가정한다면 상미분방정식

(ODE)을 이용하여 구현할 수 있다. SIR 모형에 대한 상미분방정식을 표현하기 위해서는 각 군의 상태를 전체 모집단에 대한 비율로 바꾸어 표현한다. 즉 $x(t) = X(t)/M$, $y(t) = Y(t)/M$, $z(t) = Z(t)/M$과 같이 표현되며 상미분 방정식은 다음의 식 (2.1)과 같다.

$$\begin{aligned} \frac{dx(t)}{dt} &= -\lambda x(t)y(t) \\ \frac{dy(t)}{dt} &= \lambda x(t)y(t) - \gamma y(t) \\ \frac{dz(t)}{dt} &= \gamma y(t) \end{aligned} \tag{2.1}$$

여기에서 λ와 γ는 양의 실수이다. $\lambda x(t)y(t)$는 감염 가능군과 감염군의 크기에 비례하며, 이 비율만큼 감염군이 증가한다. $\gamma y(t)$는 감염군에서 제거되어 병으로부터 회복되는 비율이며, 감염군에 비례하여 감소한다. 위에서 언급한 M으로 고정된 전체 모집단의 크기는 식 (2.1)의 모든 상미분방정식으로부터 다음과 같은 식을 추가함으로써 얻어 낼 수 있다.

$$\frac{dx(t)}{dt} + \frac{dy(t)}{dt} + \frac{dz(t)}{dt} = 0 \Rightarrow x(t) + y(t) + z(t) = M$$

상미분방정식은 페트리 넷 (petri net)에서의 화학량론 행렬 (stoichiometry matrix)을 이용하여 표현할 수 있다 (Wilkinson, 2012, page 21). 이를 유도하기 위한 과정은 아래와 같다.

step1. 반응물질 행렬 (reactants matrix; Pre)과 생산물질 행렬 (products; Post)을 구한다.

$$Pre = \begin{array}{c} \begin{matrix} S & I & R \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \end{array}, \quad Post = \begin{array}{c} \begin{matrix} S & I & R \end{matrix} \\ \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}$$

step2. 반응 행렬 (reaction matrix; A)를 구한다.

$$A = Post - Pre = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

step3. 화학량론 행렬 (stoichiometry matrix; S)를 구한다.

$$S = A^T = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} \tag{2.2}$$

step4. 상미분방정식을 행렬로 표현한다.

$$\frac{d}{dt}\begin{pmatrix} (S) \\ (I) \\ (R) \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda x(t)y(t) \\ \gamma y(t) \end{pmatrix} \tag{2.3}$$

식 (2.2)는 식 (2.1)에서 각각의 종 (species) 앞에 붙여있는 상수를 뜻하고, 그 상수를 가지고 행렬을 구현한 것이다. 식 (2.3)에서 가장 오른쪽에 있는 행렬은 2개의 반응에 대한 가중치 (weight)를 하나의 벡터로 만든 것이다. 상미분방정식은 결정적인 전염병 모형의 일종으로 상대적으로 모형이 간단해 진다

는 장점이 있다. 이러한 장점 때문에 매우 복잡한 형태의 모형도 상미분방정식을 이용하여 모형 구축을 시도할 수 있다.

질병 확산 모형을 구축하는 데 있어서 중요한 관심사 가운데 하나는 질병의 확산 여부이다. 또한 확산이 진행된다면 시간의 흐름에 따라 어디까지 확산이 될 것이며, 이후 이 확산 속도는 다시 감소 될 것이고 그 감소는 언제부터인가에 대한 것이다. 기본적으로 감염자의 수가 증가하는 추세라면 해당 질병은 확산되는 과정에 놓여있다고 볼 수 있다. 즉, 식 (2.1)에서 $dy(t)/dt > 0$이 됨을 의미하며 다음과 같은 식을 유도할 수 있다.

$$\begin{aligned} \frac{dy(t)}{dt} &> 0 \\ \lambda x(t)y(t) - \gamma y(t) &> 0 \\ \frac{\lambda x(t)y(t)}{\gamma} &> y(t) \end{aligned} \tag{2.4}$$

질병 확산의 초기 단계에서는 전체 모집단의 대부분의 사람들이 병에 감염될 가능성을 가지고 있기 때문에 $x(t) \approx 1$이라고 할 수 있다. 따라서 $x(t) = 1$을 식 (2.4)에 대입하면 다음과 같이 정리된다 (Jones, 2007).

$$\frac{\lambda}{\gamma} > 1 \tag{2.5}$$

식 (2.5)로부터 계산된 값은 $R_0 = \lambda/\gamma$로 정의한다. 이 수치는 기초감염재생산수 (basic reproduction number)라 불리며 질병 확산 모형의 구현에서 매우 중요하게 다루어져야 하는 수치이다. 이 값의 의미는 모든 인구가 질병에 감염될 수 있는 사람이라고 가정할 때, 감염성이 있는 환자가 감염 가능기간 동안 직접 감염시키는 평균 인원수로 전염병이 인구를 통해 확산 할 수 있는지 여부를 결정하는데 중요한 역할을 할 뿐만 아니라, R_0 값을 통해 바이러스의 규모와 추요한 백신 수를 추측할 수 있다.

$R_0 < 1$이면 한명의 감염자가 자신의 감염기간 동안에 평균적으로 한명 미만의 2차 감염자를 발생시키기 때문에, 시간에 따라 감염자 수가 감소하게 된다. 즉, 국소구간에서 점진적으로 안정되는 현상 (locally asymptotically stable)을 갖고 전염병은 더 이상 확산되지 않는다. 반대로 $R_0 > 1$이면 한명의 감염자가 자신의 감염기간 동안에 평균적으로 2명 이상의 2차 감염자를 발생시키기 때문에, 시간에 따라 감염자 수가 증가하게 된다. 이때는 안정되지 않으며 전염병이 확산 될 수 있음을 의미한다. 예를 들어 R_0 값이 10인 유행병이 있다면, 한 사람의 감염자에 의해 10명의 추가 감염자가 생길 수 있다. 이 경우에 총 국민의 90%이상이 백신을 맞아야 10명중 1명 이하의 비율로 감염되어 R_0가 1이하가 되고 유행을 종식시킬 수 있다. $R_0 = 1$이면 풍토병이라고 하는데, 풍토병이란 특정 지역에서 사는 주민들에서 지속적으로 발생하고 있는 전염병을 말한다. SIR 모형에서의 R_0은 λ/γ로 정의되지만, 복잡한 모형에서는 여러 다른 상수들이 R_0에 영향을 미치기도 한다 (Lee, 2011; Korea Centers for Disease Control and Prevention, 2011).

식 (2.1)의 상미분방정식에 의한 SIR 모형의 두 개의 모수 λ와 γ를 추정하기 위하여 데이터로부터 편차의 제곱합을 최소화하는 최소제곱법 (least square estimation; LSE)을 이용한다. 최소제곱법에 의하여 식 (2.6)을 최소화하는 모수의 값을 찾아 모수의 추정값으로 사용한다.

$$\sum_{t=0}^{n} (y(t) - \hat{y}(t))^2 \tag{2.6}$$

여기서 $\hat{y}(t)$은 식 (2.1)의 상미분방정식에 의해 예측된 수이다.

3. 확률적 질병 확산 모형

식 (2.1)로 표현된 질병 확산 모형은 결정적인 과정을 통하여 질병의 확산 과정을 설명하고자 하는 모형이다. 그러나 현실적으로 질병의 확산 과정이 이와 같이 결정적 형태로 움직인다고 보는 것 보다는 확률적 (stochastic)인 과정으로 움직인다고 보는 것이 보다 적절 할 수 있다 (Choi와 Rempala, 2012; Andersson과 Britton, 2000, page 3). 확률적인 과정을 가정한 질병 확산 모형은 확률적 반응 모형 (stochastic chemical reaction model, stochastic kinetic network model)을 통하여 구현할 수 있다. 확률적 반응 모형은 원래 생물학적 시스템 (biological system)의 움직임이 확률적으로 움직인다는 가정하에 모형을 구현하기 위해 개발된 모형이다. 전체 모형을 종 (species)과 반응 (reaction)을 통하여 구현할 수 있다. 고전적인 형태인 SIR 모형을 확률적 반응 모형으로 구현하면 식 (3.1)과 같다.

$$\begin{aligned} X + Y &\xrightarrow{\lambda} 2Y : h_1 = h_1(X, Y, Z) = \lambda XY \\ Y &\xrightarrow{\gamma} Z : h_2 = h_2(X, Y, Z) = \gamma Y \end{aligned} \tag{3.1}$$

식 (3.1)에서 X, Y, Z는 SIR 모형을 구성하는 세 집단으로 감염 가능군, 감염군, 회복군의 실제 집단의 크기를 나타내며, 확률 반응 모형의 각 종 (species)이 된다. 첫 번째 식과 두 번째 식은 각각 반응 (reaction)으로 불리며 전체 확률 반응 모형에서 각 종들 간의 연관을 통한 종들의 변화를 설명하기 위한 식이다. 식 (3.1)에서 첫 번째 반응식은 감염군 Y 가운데 하나의 개체가 감염 가능군 X 전체 가운데 하나의 개체와 결합하여 감염 가능군 개체 가운데 하나를 감염군으로 변화시키는 과정을 설명하는 반응식이다. 두 번째 반응식은 감염군 Y의 개체 하나가 질병으로부터 회복되어 회복군으로 이동하는 과정을 설명하는 반응식이다. 첫 번째 반응이 일어나면 X는 개체가 하나 감소하는 동시에 Y는 개체가 하나 증가한다. 두 번째 반응이 일어나게 되면 Y 개체가 하나 감소하는 동시에 Z 개체가 하나 증가하게 된다. 각각의 반응식은 반응 상수 (reaction constant)가 할당된다.

확률적 반응 모형에서 각 종들은 각각 연속 시간 마코프 연쇄 (continuous time markov chain)를 따른다고 가정하며, 각 반응의 발생은 위험함수 (hazard function) h_1, h_2에 비례하여 발생한다. 결과적으로 X와 Y의 생성은 각각의 위험함수 h_1과 h_2를 모수로 하는 포아송 과정 (poisson process)을 따른다고 볼 수 있다.

반응 상수가 주어진 경우, 이와 같은 화학 반응 모형의 구현을 위하여 다양한 방법이 제시되고 있다. 이 가운데 가장 대표적인 방법은 Gillespie (1977)에 의해 고안된 확률 시뮬레이션 알고리즘 (stochastic simulation algorithm)이다. Gillespie 알고리즘은 크게 두 가지 부분으로 구성된다. 어떠한 반응이 발생하는가에 대한 부분과 그 반응이 발생했을 때까지 시간이 얼마나 흘렀는가에 대한 부분이다. 확률적 반응 모형에서 모든 위험함수의 합을 h_0라 하자. 식 (3.1)의 확률적 SIR 모형에 대한 반응 모형에서는 $h_0 = h_1 + h_2$가 된다. 이제 특정 반응이 발생할 때까지의 시간은 h_0를 모수로 가지는 지수분포 (exponential distribution)를 따른다 가정하고, 특정 반응의 생성은 h_i/h_0를 확률로 가지는 이산 확률분포 (discrete probability distribution)를 따른다고 가정한다. 이러한 가정하에 Gillespie 알고리즘의 구현 과정은 다음과 같이 주어진다.

step1. 초기 시점에서의 각 종들의 수와 반응 상수가 주어져야한다.

step2. 현재 상태 종에 기반해서 각 시점에서의 h_i을 구한다.

step3. 각 시점에서의 h_i을 모두 더한다.
$h_0 = \sum h_i$

step4. 다음 반응이 발생했을 때까지의 시간을 추출한다. $t' \sim Exp(h_0)$.

step5. 시간을 업데이트한다. 여기서, t는 현재까지 시간, $t^{'}$는 샘플링 되는 시간이다.
$t = t + t^{'}$

step6. 어떤 반응이 발생 했는지를 추출한다.

step7. 어떤 반응이 발생했는지에 따라 종의 숫자가 달라지므로 종을 업데이트한다.

step8. 종과 시간을 출력한다.

step9. $t > T_{max}$이면 시뮬레이션을 종료하고, 그렇지 않으면 Step 2로 되돌아간다.

Figure 3.1은 식 (2.1)에서 제시한 결정적 과정을 따르는 SIR 모형과 Gillespie 알고리즘을 이용하여 구현한 식 (3.1)의 확률적 SIR 모형을 구현한 그림이다. 각 종의 초기값은 $X(0) = 500$, $Y(0) = 1$, $Z(0) = 0$이며 전체 모집단의 수는 $M = 501$로 하였다. 또한 모형 구현을 위한 반응 상수는 $\lambda = 0.5$, $\gamma = 0.1$로 하였다. 그림에서 매끈한 곡선으로 표현한 부분이 결정적 모형을 나타낸다. 가느다란 점선으로 표시된 부분은 이에 대응하는 확률적 모형을 나타낸다.

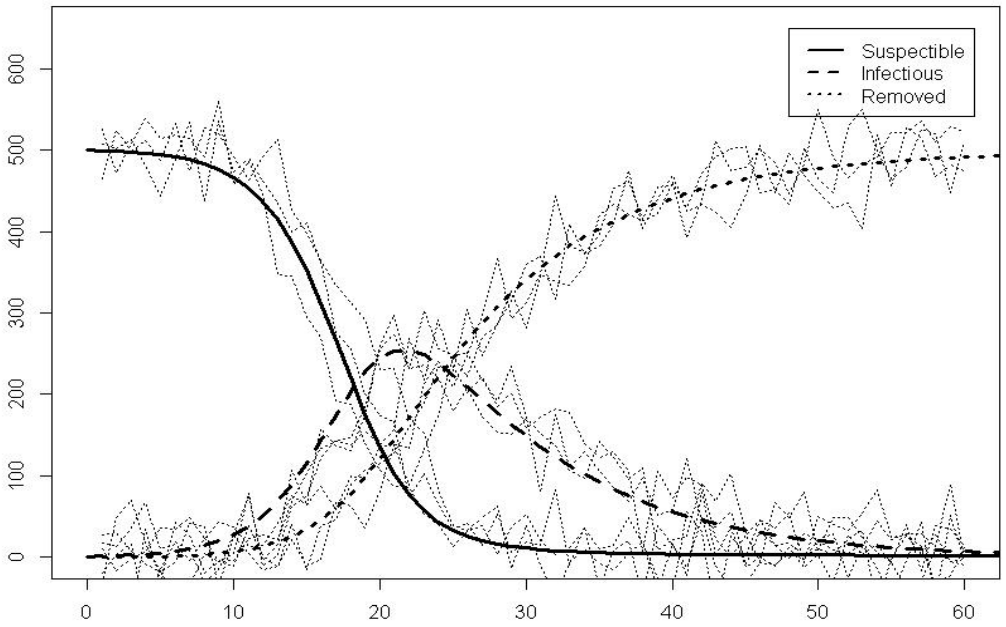

Figure 3.1 ODE and SKN trajectories for SIR model species with total population size M = 501

4. 자료분석

전염병의 확산 예측 모형을 구현하기 위한 데이터는 질병관리본부 (Korea Centers for Disease Control and Prevention)에서 관리하고 있는 감염병 웹통계 시스템의 2001년 1월부터 2002년 3월까지 (총 62주)의 국내 말라리아의 주별 발병자 수 자료를 사용하여 분석하였다. 말라리아의 경우 모든 사람이 감염 대상군이 될 수 있으며 최초 관찰시점인 2001년 1월의 인구수는 45,524,681명으로, 이를 모집단 수로 볼 수 있다. 수집된 자료로부터 최초 관찰시점에서 감염된 사람의 수는 2명으로 보고되었으며, 일 단위가 아닌 주 단위로 총 62시점의 주별 신규 감염자수가 수집되었다.

Figure 4.1의 첫번째 그림은 관찰된 주별 신규 말라리아의 감염자 수에 대한 시도표를 나타낸다. 최초 발생시점에서 시작하여 꾸준히 증가하던 감염자의 수는 30주차에서 최고치를 기록하며 이후 다시 감염자수가 감소하게 된다. 그리고 62주차 이후에는 더 이상 신규 감염자가 나타나지 않았다. 신규 감염자를 숫자만을 가지고 SIR 모형에 적합하기 위해서는 추가적인 자료의 변환이 필요하다. 실제 SIR 모형에서 각각 S, I, R로 표시되는 감염 가능군, 감염군, 회복군의 시점별 숫자가 필요하다. 즉, 전체 시점에 대한 각 군 (확률 반응 모형에서의 종)에 대한 확산 과정 (trajectory)이 필요하게 된다. 하지만 관찰된 데이터는 오직 신규로 감염된 사람만의 숫자가 관찰되어 있다. 이제 주어진 정보를 이용하여 먼저 전체 확산 과정에 따른 자료 구축 과정에 대하여 알아보자.

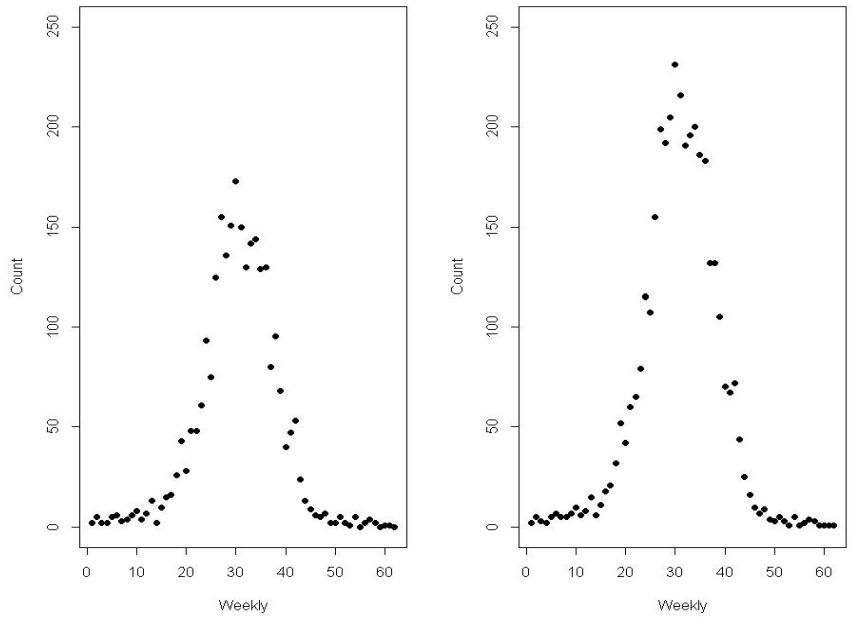

Figure 4.1 Weekly count of infected. Left panel is the daily reported number of newly infected people and right panel is the trajectory of whole stage of Infected species $Y(t)$ in the SIR model.

첫 번째로 전체 모집단의 숫자와 신규 감염자의 숫자를 알고 있으므로 이를 통하여 감염 가능군의 경로를 구축할 수 있다. 최소 시점에서부터 각 시점에서 발생한 신규 감염자의 숫자를 누적해서 빼 줌으로 감염 가능군의 경로를 구할 수 있다. 예를 들어 $X(t)$를 시점 t에서 감염 가능군의 수라 하고 r_{1t}를 시점 t에서 질병에 감염된 사람의 숫자라 하자. Figure 4.1의 첫 번째 그림이 이 r_{1t}에 대한 시도표라고 할 수 있다. 이제 $X(t)$는 다음과 같이 계산된다.

$$X(t) = X(t-1) - r_{1t}$$

결과적으로 r_{1t}는 식 (3.1)의 확률반응모형에서 첫번째 반응 (reaction)의 발생 빈도를 나타낸다.

다음으로 감염군 $Y(t)$의 구축과정을 알아보자. $Y(t)$는 결국 시점 t에서 여전히 질병에 감염된 채로 남아있는 숫자에 신규로 감염된 사람의 숫자를 더하고 질병으로부터 회복되어 회복군으로 이동한 숫자를 차감하여 구할 수 있다. r_{2t}를 시점 t에서 회복된 사람의 숫자라 하면 감염군의 확산 과정 다음과 같이 계산된다.

$$Y(t) = Y(t-1) + r_{1t} - r_{2t} \tag{4.1}$$

감염군의 확산 과정을 구축하기 위해서는 관찰 시점에서 회복군으로 이동한 숫자인 r_{2t}를 알아야 한다. 이 r_{2t}는 식 (3.1)의 확률 반응 모형에서 두 번째 반응 (reaction)의 발생숫자이다. 기존의 문헌 연구를 통하여 우리나라에서 말라리아의 경우 평균 3일의 회복기간이 필요하게 된다 (Yeom과 Park, 2007).

본 연구에서는 3일을 회복기간으로 하여 감염군의 확산 과정을 구축하였다. 먼저 관찰된 자료는 주별 자료이기 때문에 이를 먼저 고려한다. 3일의 회복기간을 고려하면 7일 가운데 최초 5일 이내에 질병이 발생한 경우 다음 관찰시점으로 넘어가기 이전에 질병으로부터 회복 된다고 할 수 있다. 그리고 마지막 2일 동안 질병에 걸린 사람은 다음 시점까지 질병에 감염 상태로 이동하게 된다. 이를 고려하게 되면 $t-1$ 시점에서 새로 질병에 감염된 사람 가운데 평균적으로 5/7는 $t-1$ 시점 내에서 질병으로 회복되고 나머지 2/7의 사람만이 다음 시점인 t 시점까지 감염군에 머무르게 된다. 즉, 식 (4.1)에서 $Y(t-1)$에 해당하는 숫자는 이전 시점의 전체 감염군의 숫자가 아니고 $t-1$ 시점에서 신규로 질병에 감염된 사람 가운데 2/7에 해당하는 숫자가 된다. 그리고 r_{2t}는 유사한 가정을 적용하여 $Y(t-1)$과 $r_{1t} \times (5/7)$의 합으로 계산된다. 모든 숫자는 소수점으로 계산되는 경우 소수점 아래 첫번째 자리에서 반올림 하였다.

마지막으로 회복군의 전체 경로를 계산한다. 전체 모집단이 고정되어 있다고 가정하였기 때문에 감염가능군과 감염군의 숫자가 결정이 되면 회복군의 숫자는 다음의 식에 따라 쉽게 구할 수 있다.

$$Z(t) = M - X(t) - Y(t)$$

이와 같은 과정을 수행하여 관찰된 데이터와 적절한 가정을 통하여 감염 가능군, 감염군, 회복군에 대한 전체 확산 과정을 구축할 수 있다. Figure 4.1의 두 번째 그림은 이와 같은 과정을 통하여 재구축된 감염군의 확산 과정을 나타내는 그림이다. 이제 구축된 전체 확산 과정을 결정적인 과정에 따라 확산한다고 보고 주어진 자료를 이용하여 모형 적합을 수행하였다.

식 (2.1)의 상미분 방정식에 의한 SIR 모형의 모수를 추정하기 위하여 식 (2.6)을 이용하였다. 상미분 방정식을 위한 초기값으로 $x(0)$, $y(0)$, $z(0)$를 할당하여야 한다. 관찰된 자료를 이용하여 초기값을 할당할 수 있는데 최초 시점에서 감염 가능군과 감염군의 숫자를 각각 n, m 이라 하면, 즉 $X(0) = n$, $Y(0) = m$이 된다. $Z(0) = 0$이라 하면 쉽게 $n + m = M$이 된다. 관찰된 자료에 의하여 $m = 2$가 되고 $n = M - m = 45,524,681 - 2 = 45,524,679$가 된다. 상미분방정식을 위한 초기값 $x(0)$와 $y(0)$는 다음과 같이 계산된다.

$$x(0) = 1 - \frac{m}{M}, \quad y(0) = \frac{m}{M} \tag{4.2}$$

식 (4.2)로부터 $x(0)$와 $y(0)$는 각각 0.999999956과 0.000000043으로 계산된다. 모형 구축을 위하여 필요한 모수는 감염률과 회복률을 나타내는 모수인 λ와 γ이다. 여기서 회복률를 나타내는 γ는 회복기간의 역수 관계가 있다. 전술한 바와 같이 분석에 사용된 말라리아의 경우 평균 회복기간이 3일로 이를 주로 바꾼 후 역수를 취하면 2.33이 된다. 따라서 $\gamma = 2.33$으로 고정한 후 최소제곱법 (LSE)을 이용하여 감염률를 나타내는 λ를 추정하고자 하였다. 추정결과 2.3375로 계산되었다. 이로부터 계산된 R_0는 1.0032로 말라리아 질병의 확산은 일정 정도 확산 과정을 거친 후 다시 안정적으로 감소하고 있다고 볼 수 있다.

이제 추정된 두 모수를 이용하여 결정적 모형과 확률적 모형을 이용하여 모형 적합을 시도해 보자. 결정적 모형의 경우 식 (2.1)의 상미분방정식에 추정된 모수와 초기값을 적용하여 모형을 구축하였다. 확률적 모형의 경우 식 (3.1)의 확률 반응 모형을 이용하였다. 결정적 모형과 같이 추정된 모수를 적용하였으며 3장에서 설명한 Gillespie 알고리즘을 이용하여 모형을 구축하였다.

Figure 4.2는 구축된 모형 가운데 감염군의 확산 과정만을 표시한 그림이다. 가운데 굵은 곡선으로 표시된 부분이 상미분방정식을 이용하여 모형 적합을 시도한 결과이다. 확률적 모형의 경우 Gillespie 알고리즘을 100번 수행하여 총 100개의 모형을 적합하였다. 결정적 모형을 중심으로 하여 가느다란 점

선으로 표시된 부분이 확률적 모형의 구축 결과 가운데 10개를 임의로 선택하여 표시한 것이다. 또한 95% 신뢰구간의 상한과 하한을 함께 표시하였다. Figure 4.2에서 검은 점으로 표시된 부분은 실제 자료의 감염군을 나타낸다. 구축된 모형이 실제 데이터를 정확히 설명하고 있지는 않으나 확산 과정의 최고점을 중심으로 보았을 때는 잘 예측한다고 볼 수 있다.

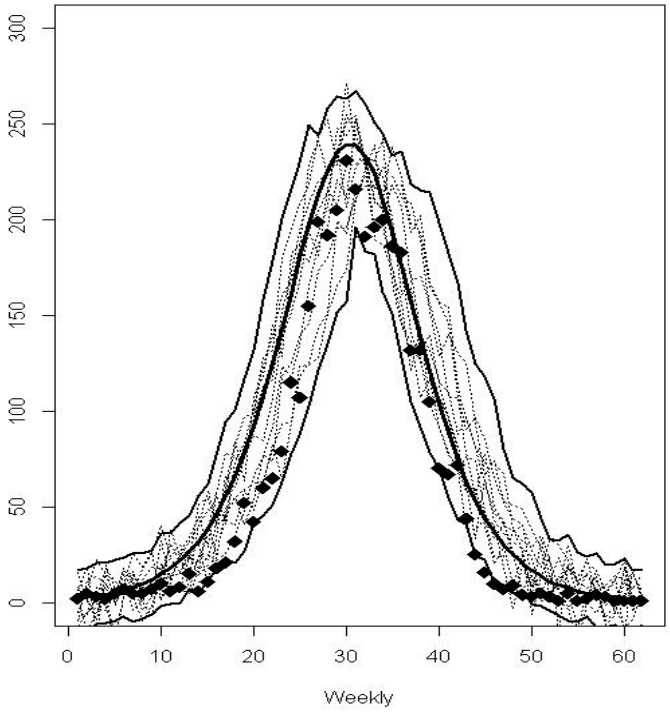

Figure 4.2 Model fitting results for infected

5. 결론

본 연구에서는 질병의 확산 과정을 설명하고자 하는 수리적으로 표현한 전염병 모형을 구축하고자 하였다. 질병 확산 모형으로는 고전적인 형태인 SIR 모형을 이용하였다. 특히 질병의 확산 과정이 결정적인 과정과 더불어 확률적인 과정에 따라 확산한다고 보고 모형을 함께 구축해 보고자 하였다. 결정적 모형과 확률적 모형간의 비교를 수행하였다. 모수의 추정을 위해서는 최소제곱법을 이용하여 모수를 추정하였으며 추정된 모수를 이용하여 결정적 모형과 확률적 모형을 구축하여 보았다. 모수의 추정에 있어서는 결정적 모형만을 고려한 후 모수를 추정하였다. 그러나 진정한 확률적 모형을 구축하기 위해서는 모형의 추정 단계에서부터 확률적 모형을 가정하고 모형의 추정과정을 진행하여야 할 것이다. 본 연구의 추후 과제로는 확률적 모형을 기반으로 하여 모수의 추정을 수행한 후 추정된 결과를 바탕으로 하는 모형 구축을 시도해 볼 수 있을 것이다. 또 다른 측면으로는 질병 확산 모형 가운데 보다 복잡 하지만 실제 현상을 더 적절하게 반영할 수 있는 확장된 질병 확산 모형을 적용하여 모형을 구축을 시도한 후 그 결과들에 대한 상호 비교 과정도 진행 할 수 있을 것이다.

References

Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*, Springer, New York.

Bernoulli, D. (1766). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. *Histoire de l'Académie Royale des Sciences: Mémoires de Mathématiques et de Physique*, 1-40.

Choi, B. and Rempala, G. A. (2012). Inference for discretely observed stochastic kinetic networks with applications to epidemic modeling. *Biostatistics*, **13**, 153-165.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, **81**, 2340-2361.

Graunt, J. (1662). Natural and political observations made upon the bills of mortality. `http://www.neonatology.org/pdf/graunt.pdf`.

Hamer, W. H. (1906). *The Milroy lectures on epidemic disease in England*, Bedford Press, England.

Hwang, J. H. and Oh, C. H. (2014). A study on the spread of the foot-and-mouth disease in Korea in 2010/2011. *Journal of the Korean Data & Information Science Society*, **25**, 271-280.

Hwang, N. A., Jeong, B. Y., Lim, Y. C. and Park, J. S. (2007). Diseases data analysis using sir nonlinear regression model. *Journal of The Korean Data Analysis Society*, **9**, 49-59.

Jones, J. H. (2007). *Notes on R_0*, Department of Anthropological Sciences Stanford University, California.

Kim, E. Y., Lee, S., Byun, Y. T., Lee, H. J. and Lee, T. J. (2013). Implementation of integrated monitoring system for trace and path prediction of infectious disease. *Journal of Korean Society for Internet Information*, **14**, 69-76.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A: Mathematical Physical and Engineering Sciences*, **115**, 700-721.

Korea Centers for Disease Control and Prevention. (2011). *Developing a mathematical model of tuberculosis and proposing optimal strategies for preventing tuberculosis in South Korea*, Report, Korea Centers for Disease Control and Prevention, Chungbuk.

Lee, J. H., Murshed, M. S. and Park, J. S. (2009). Estimation of infection distribution and prevalence number of Tsutsugamushi fever in Korea. *Journal of the Korean Data & Information Science Society*, **20**, 149-158.

Lee, S. G., Ko, R. Y. and Lee, J. H. (2010). Mathematical modelling of the h1n1 influenza. *Journal of the Korean Society of Mathematical Education Series E*, **24**, 887-889.

Lee, S. W. (2011). *Study on the chronological analysis and follow-up model of epidemic models*, Master Thesis, Korea University, Seoul.

Ross, R. (1911). *The prevention of malaria*, 2nd Eds., John Murray, London.

Wilkinson, D. J. (2012). *Stochastic modelling for systems Biology*, 2nd Eds. CRC Press, Boca Raton.

Yeom, J. S. and Park, Y. K. (2007). Treatment of Korean vivax malaria in Korea. *Journal of the Korean Medical Association*, **50**, 88-92.

Journal of the Korean Data & Information Science Society
2015, **26**(2), 301–312

http://dx.doi.org/10.7465/jkdi.2015.26.2.301
한국데이터정보과학회지

Development of epidemic model using the stochastic method[†]

Soorack Ryu[1] · Boseung Choi[2]

[1]Department of Statistics, Daegu University
[2]Department of Statistics and Computer Science, Daegu University

Received 5 January 2015, revised 29 January 2015, accepted 10 February 2015

Abstract

The purpose of this paper is to establish the epidemic model to explain the process of disease spread. The process of disease spread can be classified into two types: deterministic process and stochastic process. Most studies supposed that the process follows the deterministic process and established the model using the ordinary differential equation. In this article, we try to build the disease spread prediction model based on the SIR (Suspectible – Infectious – Recovered) model. we first estimated the model parameters using least squared method and applied to a deterministic model using ordinary differential equation. we also applied to a stochastic model based on Gillespie algorithm. The methods introduced in this paper are applied to the data on the number of cases of malaria every week from January 2001 to March 2003, released by Korea Centers for Disease Control and Prevention. As a result, we conclude that our model explains well the process of disease spread.

Keywords: Epidemic model, Gillespie algorithm, ODE model, SIR model, stochastic kinetic network model.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No.2012R1A1A1010156).
[1] Graduate student, Department of Statistics, Daegu University, Gyeongbuk 712-714, Korea.
[2] Corresponding author: Assistant professor, Department of Statistics and Computer Science, Daegu University, Gyeongbuk 712-714, Korea. E-mail: bchoi@daegu.ac.kr