

# 일반화선형혼합모형을 통한 해충밀도의 시공간분포 연구

박홍선<sup>1)</sup> 조기종<sup>2)</sup>

## 요약

농작물에 기주하는 해충의 밀도추정은 해충종합적 관리(Integrated Pest Management: IPM)의 중요한 연구분야이다. 이는 농약을 살포한다든지 천적을 방사하는 것과 같은 인위적인 방제관리의 성패여부가 해충밀도에 대한 정보에 의존하기 때문이다. 본 연구는 온실장미에서 서식하는 점박이용애의 밀도분포를 일반화 선형혼합모형을 통하여 분석하고 있다. 자료습득이 식물체의 일정 부위에서 반복적으로 이루어 졌기 때문에 경시적 자료분석의 개체특정(Subject-Specific: SS)방법과 모집단평균(Population-Averaged: PA)방법을 모두 사용한 후에, 그 결과를 비교, 검토 하였다.

*Keywords:* GEE, GLIMMIX, 반복측정자료, GLMM

## 1. 서론

온실을 이용한 장미 재배사업은 국내의 대표적인 화훼사업 중의 하나이면서 사업의 경제적 규모도 증대되고 있기 때문에, 온실장미에 대한 해충관리 또한 그 중요성이 더해가는 현실이다. 온실장미에 대한 대표적인 해충으로 점박이용애(Two-Spotted Spider Mites: TSSM)를 들 수 있는데, 점박이용애는 장미 잎 뒷면에 서식하면서 잎의 엽록소를 흡즙함으로써 장미의 미적 가치는 물론 장미의 정상적인 성장에 커다란 저해요소가 되고 있다. 성충의 크기는 0.3-0.5 밀리미터 밖에 안 되어 육안으로 식별이 어려울 뿐 아니라, 거미처럼 줄을 치며 이동을 하는 습성이 있고, 섭씨 25도를 기준으로 약 30일의 수명을 가지고 있는 것으로 알려져 있다.

그림1.1은 점박이용애에 의해 피해입은 장미 잎과 이들 뒷면에 서식하는 점박이용애들의 모습이며, 그림1.2은 점박이용애를 확대 촬영한 모습이다. 온실장미가 점박이용애로부터 피해를 입게 되면, 노란 반점이 잎 표면에 생기기 시작하고, 이러한 증상이 심해지면 잎의 색깔이 탈색되는 현상까지 발생하게 된다. 이 해충의 피해를 막기 위해서, 시간에 따른 해충의 밀도변화나, 식물체 내에 해충의 분포형태를 조사할 필요가 있는데, 이는 화학적 방제나 생물학적 방제를 사용할 때 방제 시기나 방제 위치 결정에 대한 과학적 근거로 활용될 수 있기 때문이다.

1) (449-791) 경기도 용인시 모현면 왕산리 산 79 한국외국어대학교 통계학과 교수

E-mail: hspark@hufs.ac.kr

2) (136-701) 서울시 성북구 안암동 5가 1번지 고려대학교 환경생태공학부 부교수

E-mail: kjcho@korea.ac.kr



그림 1.1: 장미 잎사귀 뒷면에 서식하는 점박이응애



그림 1.2: 점박이 응애를 확대한 사진

이 연구를 위하여 우리는 부여에 있는 장미 온실에 24 개의 장미나무를 임의로 선택하고, 장미나무 안에 상/중/하의 수직적 위치를 설정하고, 각 위치에 대해 3개의 잎(1개 잎은 5개의 소엽으로 구성됨)을 채취하여 큰 9개의 엽 위에 있는 점박이응애 성충의 수를 조사하였다. 간단히 말하면, 상/중/하, 각 위치당 9개의 엽들을 채집한 것인데 표본위의 움직임은 응애들을 고정시키기 위해 우리는 3M 스프레이 접착제를 사용하였고, 채집된 표본들은 실험실에서 계수작업을 하였다. 이와 같은 표본수집 과정은 1997년 6월부터 11월까지 한달에 한번씩 이루어 졌다.

본 연구의 목적은 점박이응애의 밀도가 식물내의 위치(상/중/하)에 따라 차이가 있는지, 차이가 있다면 어느 위치에 해충 밀도가 가장 높은지에 대한 조사와, 시간에 따르는 해충 밀도의 변화추이를 조사하는데 있었다. 이를 위해, 우리는 Liang 과 Zeger (1986)가 제안한 반복측정자료에서의 일반화 추정방정식(Generalized Estimating Equation : GEE)을 사용했으며, 이 방법은 개체가 속한 모집단에 대한 추론에 초점을 맞추었다고 하여서 모집단평

균(Population-Averaged : PA) 접근 방법(Zeger, Liang 그리고 Albert, 1988)이라고 일컫고 있다. 이 외에 개체특성(Subject-Specific : SS) 접근방법의 하나인 벌점준가능도(Penalized Quasi-likelihood: PQL) 추정방법(Breslow 와 Clayton, 1993)을 사용했는데, 이 방법은 일반화 선형모형(McCullagh 와 Nelder, 1989)의 기본틀에 임의효과를 첨부시킨 알고리즘으로써 Wolfinger 와 O'connell (1993)이 작성한 GLIMMIX 라는 매크로 프로그램(SAS, 1992)이 채택한 방법이다. 이 SS 접근방법은 PA 접근방법과는 달리, 각 개체의 프로파일에 초점을 맞추므로써 전체 모집단을 해석하고 있다.

## 2. GEE 와 PQL

먼저, PA 접근방법의 하나인 GEE 에 대해 설명해 보도록 하겠다. 확률변수  $y_{ij}$  가  $i$  번째 개체의  $j$  번째 시점의 관측값을 나타낸다고 할 때, 벡터  $Y_i = (y_{i1}, y_{i2}, \dots, y_{it_i})^T$  는 개체  $i$  에 대한  $t_i \times 1$  ( $t_i \leq t$ ) 인 반응벡터가 되고,  $X_i = (x_{i1}, x_{i2}, \dots, x_{ip})$  는  $t_i \times p$  인 공변량행렬이 된다. 여기서  $x_{ij}$  는  $t_i \times 1$  인 벡터이다. 이때,  $y_{ij}$  의 주변 확률밀도함수는 지수족함수에 속한다는 가정하에,

$$f(y_{ij}) = \exp \left[ \frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi) \right] \quad (2.1)$$

이고,  $y_{ij}$  의 평균과 분산은 다음과 같다.

$$E(y_{ij}) = \mu_{ij} = b'(\theta_{ij}), \quad \text{var}(y_{ij}) = V(\mu_{ij})a(\phi) \quad (2.2)$$

여기서  $V(\mu_{ij}) = b''(\theta_{ij})$  는 분산함수이고  $\phi$  는 미지의 척도 모수이다.  $\eta_i = X_i\beta$  라고 놓았을 때,  $\eta_i = (\eta_{i1}, \dots, \eta_{it_i})^T$ , 그리고  $\eta_{ij} = g(\mu_{ij}) = x_{ij}\beta_j$  가 되며, 여기서  $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$  는  $p \times 1$  미지 모수의 벡터이고  $g(\cdot)$  은 연결함수이다.

포아송분포 같은 계수적 반응변수에서는,  $g(\mu_{ij}) = \log(\mu_{ij})$ ,  $V(\mu_{ij}) = \mu_{ij}$ , 그리고  $a(\phi) = 1$ 이다. 만약  $S_i = Y_i - \mu_i$ ,  $\mu_i = (\mu_{i1}, \dots, \mu_{it_i})^T$ , 그리고  $D_i = \partial\eta_i/\partial\beta$  라면,  $i$  번째 개체에 대한 준가능도(Quasi-Likelihood) 추정방정식은

$$D_i^T V_i^{-1} S_i = 0 \quad (2.3)$$

이며,  $V_i$  는  $\text{cov}(Y_i)$ 와 같다. 그러므로,  $n$  개의 독립된 개체에 대해 반복측정된 자료의 일반화 추정방정식은

$$\sum_{i=1}^n D_i^T V_i^{-1} S_i = 0 \quad (2.4)$$

가 되는데  $V_i$  는  $t_i \times t_i$  크기의 가공분산행렬(working covariance matrix)을 나타내며,

$$V_i = (A_i \Phi_i)^{1/2} R_i(\gamma) (A_i \Phi_i)^{1/2} \quad (2.5)$$

표 2.1: 대표적인  $t \times t$  가상관행렬의 구조

구조	이름	모수의 개수	행렬의 $(i, j)$ 째 원소
UN	Unstructured	$t(t+1)/2$	$\rho_{ij}I(i \neq j) + I(i = j)$
CS	Compound Symmetry	2	$\rho I(i \neq j) + I(i = j)$
AR(1)	Autoregressive(1)	2	$\rho^{ i-j }$
TOEP	Toeplitz	$t$	$\rho_{ i-j +1}, (\rho_1 = 1)$
TOEP(2)	Banded Toeplitz	2	$\rho_{ i-j +1}I( i-j  < 2), (\rho_1 = 1)$

라고 분할 표현된다. 여기서,  $A_i\Phi_i$ 는  $t_i \times t_i$  크기의 대각행렬로서  $\text{var}(Y_i) = A_i\Phi_i$  를 만족시킨다. 따라서, 가공분산행렬을 정의하기 위해서는 분산에 해당되는  $A_i\Phi_i$  와 가상관행렬(working correlation matrix)인  $R_i(\gamma)$ 가 미리 정의되어야 한다. GEE 에서는 여러 개의 가상관행렬을 시도해 봄으로써 모수 추정을 하고 있다.

표 2.1은 대표적인 가상관행렬의 종류를 나열한 것인데, 가상관행렬의 이름, 모수의 개수, 그리고 가상관행렬의  $(i, j)$ 번째 원소를 보여주고 있다.

SS 접근방법의 하나인 별점준가능도(PQL) 추정법은 임의효과  $\mathbf{u}$ 가 주어졌을 때의 조건부 기대값을  $\boldsymbol{\mu}_i = E(\mathbf{y}_i|\mathbf{u})$  라고 가정하여, 평균과 공변량과의 관계를

$$\boldsymbol{\eta}_i = g(\boldsymbol{\mu}_i) = X_i\boldsymbol{\beta} + Z_i\mathbf{u} \quad (2.6)$$

라고 정의하는데,  $\boldsymbol{\beta}$ 는  $X_i$ 라는 공변량행렬과 연관되는 미지의 고정효과 벡터이며,  $\mathbf{u}$ 는 행렬  $Z_i$ 와 대응하는 미지의 임의효과 벡터로서 평균은 0 이고 분산은  $D$ 라는 정규분포를 갖고 있다고 가정한다.

PQL 추정방법은,  $g(\mathbf{y}_i)$ 를  $\boldsymbol{\mu}_i$ 에서 Taylor 전개한  $\tilde{\mathbf{y}}_i$ 라는 가변수(pseudo-response)를 생성시키는데,  $\Delta = \text{diag}\{g_{\boldsymbol{\mu}}(\boldsymbol{\mu}_i)\}$  일 때

$$\tilde{\mathbf{y}}_i = X_i\boldsymbol{\beta} + Z_i\mathbf{u} + \Delta(\mathbf{y}_i - \boldsymbol{\mu}_i), \quad (2.7)$$

가 되어, 만일  $\Delta = \text{diag}\{g_{\boldsymbol{\mu}}(\boldsymbol{\mu}_i)\}$  대신  $\Delta^* = \text{diag}\{g_{\boldsymbol{\mu}}[g^{-1}(X_i\boldsymbol{\beta})]\}$ 로 근사시키면

$$\begin{aligned} \tilde{\mathbf{y}}_i &= X_i\boldsymbol{\beta} + Z_i\mathbf{u} + \Delta^*(\mathbf{y}_i - \boldsymbol{\mu}_i), \\ &= X_i\boldsymbol{\beta} + Z_i\mathbf{u} + \mathbf{e}^*_i \end{aligned} \quad (2.8)$$

가 되며,  $\mathbf{e}^*_i$ 는 평균이 0이고 분산이  $R = \text{cov}(\mathbf{e}^*_i|\boldsymbol{\mu}_i)$  인 오차항의 모습을 따르게 된다. 따라서, 식(2.8)로부터  $\tilde{\mathbf{y}}_i$ 는

$$\begin{aligned} E(\tilde{\mathbf{y}}_i) &= X_i\boldsymbol{\beta}, \\ \text{Var}(\tilde{\mathbf{y}}_i) &= Z_i D Z_i^T + R \end{aligned} \quad (2.9)$$

인 선형혼합모형(Linear Mixed Model)식을 이루게 된다. Brewlow 와 Clayton(1993), 그리

고 Wolfinger 와 O'Connell(1993)은 식(2.8)로 부터의 추정치  $\hat{\beta}$  와  $\hat{u}$ 가

$$\begin{aligned} H &= \begin{bmatrix} X^T W^{1/2} \hat{R}^{-1} W^{1/2} X & X^T W^{1/2} \hat{R}^{-1} W^{1/2} Z \\ Z^T W^{1/2} \hat{R}^{-1} W^{1/2} X & Z^T W^{1/2} \hat{R}^{-1} W^{1/2} Z + \hat{D}^{-1} \end{bmatrix}, \\ W &= [\Delta^*]^{-1} \\ \hat{y} &= X_i \beta + Z_i u + W^{-1}(y - \mu), \end{aligned} \tag{2.10}$$

을 가정할 때,

$$H \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T W^{1/2} \hat{R}^{-1} W^{1/2} \hat{y} \\ Z^T W^{1/2} \hat{R}^{-1} W^{1/2} \hat{y} \end{bmatrix} \tag{2.11}$$

라는 혼합모형방정식을 만족시킴을 보였다. 여기서,  $\hat{D}$  와  $\hat{R}$ 은 Newton-Raphson 이나 EM 알고리즘같은 수치해석을 통해 얻게 되는데 이 값들을 식(2.11)에 대입하면

$$\begin{aligned} \hat{\beta} &= (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} \hat{y} \\ \hat{u} &= \hat{D} Z^T \hat{V}^{-1} (\hat{y} - X \hat{\beta}) \end{aligned} \tag{2.12}$$

라는 추정값을 얻게 된다.

### 3. 자료분석

점박이응애에 대한 PA 방법을 실시하기 위하여, SAS 의 PROC GENMOD 를 부록에 있는 코드를 사용하여 분석해 보았다. 여기서, 24개의 장미나무와 나무내 위치(상/중/하)를 CLASS 변수로 설정하였고, 시간(6, 7, 8, 9, 10, 11)을 나타내는 변수 time의 제곱값을 time2, 3차 제곱값을 time3 라고 하였다. 이는 그림3.1에서 알 수 있듯이 시간의 변화에 따른 해충밀도의 변화가 2차 혹은 3차 곡선을 나타내기 때문이다. 또한 GENMOD 를 사용함에 있어, 분포를 포아송(poisson)으로 설정함으로써 링크함수를 로그(log)선택하게 되고, 개체를 sub=plant\*position 으로 정의함으로써, 반복측정이 장미나무안의 상/중/하 위치에서 이루어 진다는 것을 나타내고 있다. 문제는, 상관행렬의 설정인데, GEE 에는 오직 관측치의 평균과 분산만을 가정하고 있어서, 여러가지 가능한 가상관행렬들을 시도하길 권하고 있다. 본 연구에서는 이러한 가상관행렬 중에 가장 많이 사용되는 UN, CS, AR(1), TOEP, TOEP(2) 를 시도해 보았다. 표3.1은 GEE 분석 결과를 나타내고 있는데, 가상관행렬에 관계없이 장미나무의 상층부(up)의 밀도가 중층부(mid)나 하층부(down) 밀도보다 현저히 높은 것을 알 수 있다. 예를 들면, 가상관행렬이 CS 인 경우, 상층부가 중층부에 비하여 그 밀도가  $\exp(0.4) = 1.49$  배 이고, 하층부에 비해서  $\exp(0.37) = 1.45$  배라고 할 수 있다.

또한 중층부와 하층부의 차이를 검정하기 위해 contrast 문을 사용하였는데(참조: 부록), 유의확률이 0.81-0.91 정도로써 두 위치간의 차이는 유의하지 않았다.

그림3.1은 PROC GENMOD 에서 구한 점박이응애의 평균밀도 추정치를 시간에 대해 그려본 것이다. 상층부의 밀도가 중층부나 하층부의 밀도보다 높게 표현되고 있다. 이러한

표 3.1: GEE 방법을 사용한 온실장미 내 점박이응애 평균밀도에 대한 모수추정값(괄호안은 표준오차)

모수	UN	CS	AR(1)	TOEP	TOEP(2)
Intercept	1.37 (0.61)	1.27 (0.60)	1.29 (0.62)	1.31 (0.62)	1.34 (0.63)
down	-0.35 (0.18)	-0.37 (0.18)	-0.38 (0.18)	-0.37 (0.18)	-0.36 (0.18)
mid	-0.36 (0.16)	-0.40 (0.15)	-0.40 (0.15)	-0.40 (0.15)	-0.40 (0.15)
up	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
time	2.51 (0.78)	2.62 (0.74)	2.59 (0.76)	2.57 (0.77)	2.53 (0.78)
time2	-1.18 (0.26)	-1.20 (0.25)	-1.19 (0.26)	-1.18 (0.26)	-1.18 (0.27)
time3	0.13 (0.03)	0.13 (0.02)	0.13 (0.03)	0.13 (0.03)	0.13 (0.03)

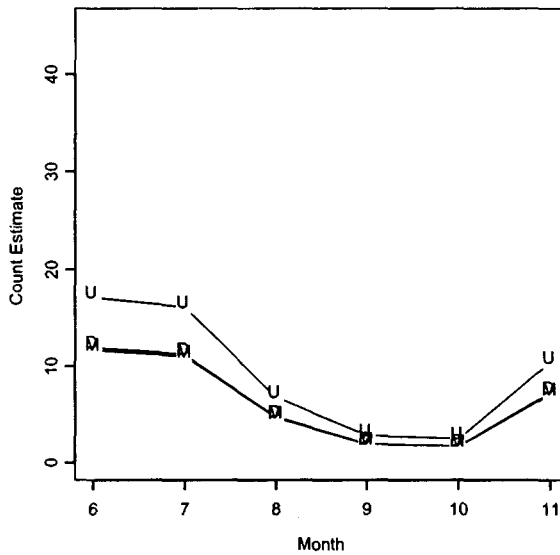


그림 3.1: GEE 에 의한 점박이응애 평균밀도 추정치: U=상층부, M=중층부, D=하층부

현상은 대체로 상층부의 잎이 연하고 부드러운 응애들이 선호하기 때문이라고 추측된다. 또한, 시간에 흐름에 따르는 점박이응애의 밀도 변화추이를 살펴보면, 6월-7월에 가장 높은 밀도를 나타내다가 8월-10월에 낮은 밀도를 보여주고 있는데, 그 이유는, 8월-10월이 장미의 비수기에 해당되어 온실에 대한 관리가 소홀하기 때문으로 추정되는데, 온실에 대한 관리소홀은 곧 식물및 해충의 서식여건의 악화로 이어지기 때문이다. 뿐만아니라, 이 시기에는 온실내부의 온도가 너무 높아져서 점박이응애의 발육률과 성장률이 급격히 떨어지기 때문이라고 할 수 있다.

같은 자료를 가지고, 이번에는 개체특정(SS) 접근방법인 GLIMMIX 매크로 프로그램으로 분석해 보았다. 부록은 이에 대한 코드를 나타내 주고 있는데, PROC GENMOD 와 다른 점은, GLIMMIX 경우에 반복측정에 의한 오차의 공분산 외에 또 다른 임의효과를 선정해 줄 수 있다는 것이다. 점박이응애의 경우, 장미나무에 해당되는 plant 는 성격상 임의효과로 간주함이 옳다고 판정되며, 그 결과 같은 나무에서 관측된 상/중/하의 관측치는 서로 양의 상관관계에 존재하게 된다. 이와 같은 성질은 매우 바람직한 것으로, 점박이응애의 생태적 특징상, 만약 어떤 나무의 한 부분에 대한 점박이응애의 밀도가 높게 관측된다면 이는 같은 나무의 다른 부분에서도 밀도가 높을 가능성이 있기 때문이다.

표 3.2: 공분산행렬에 대한 적합도 추정: -2(REML 의 로그변환 값), AIC, BIC, 그리고 Scaled Deviance 값을 나타낸 것으로, 값이 작을 수록 적합도는 우수함.

기준	UN	CS	AR(1)	TOEP	TOEP(2)
-2 Res Log L	1353	1349	1340	1335	1337
AIC	1397	1335	1346	1349	1343
BIC	1423	1358	1350	1357	1347
Sc.Dev	4346	342	332.9	335.6	331.6

GLIMMIX 매크로가 가지는 또 다른 잇점은 상관행렬에 대한 적합도를 AIC(Akaike's Information Criterion) 나 BIC(Bayesian Information Criterion)로 측정할 수 있다는 것이다. 표3.2는 GLIMMIX 의 여러 가상관행렬에 대한 이 자료의 적합도를 나타낸 것이다. 여기서 Res Log L은 제한우도(restricted likelihood)에 로그변환한 값을 의미하고 Sc.Dev는 척도화된 이탈도(Scaled Deviance)를 의미하는데, 이 값들이 작을 수록 좋은 적합도를 의미한다.

따라서, 상관행렬은 TOEP(2) 가 가장 적합한 것으로 추정되며, TOEP(2)의 모습은

$$\begin{pmatrix} 1 & \rho_1 & 0 & \cdots & 0 \\ \rho_1 & 1 & \rho_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \rho_1 & 1 \end{pmatrix} \quad (3.1)$$

로서, 반복측정에 의한 상관관계가 이웃하는 시점까지만 존재하는 형태를 말해 주고 있다. 이는 아마도 점박이응애의 생리주기가 25-30일 이고, 관측이 대략 한달에 한번씩 이루어 졌기 때문에 이웃하는 시점까지만 강한 양의 상관관계를 나타내고 있는 것으로 추측된다.

표 3.3: GLIMMIX 방법을 사용한 온실장미 내 점박이응애 평균밀도에 대한 모수추정값

모수	추정치	표준오차	t	p-value
Intercept	1.30	0.52	2.47	0.02
down	-0.37	0.17	-2.19	0.03
mid	-0.38	0.17	-2.31	0.02
up	0.00	.	.	.
time	2.58	0.63	4.11	<0.01
time2	-1.20	0.22	-5.54	<0.01
time3	0.14	0.02	6.43	<0.01

표 3.4: 온실장미내 위치에 따른 점박이응애 밀도차이에 대한 GLIMMIX 의 대조검정(Estimate 문 사용)

Label	Estimate	S.E.	DF	t	Pr>  t
pos mid to up	-0.38	0.17	348	-2.31	0.02
pos down to up	-0.37	0.17	348	-2.19	0.03
pos down to mid	0.01	0.19	348	0.05	0.96

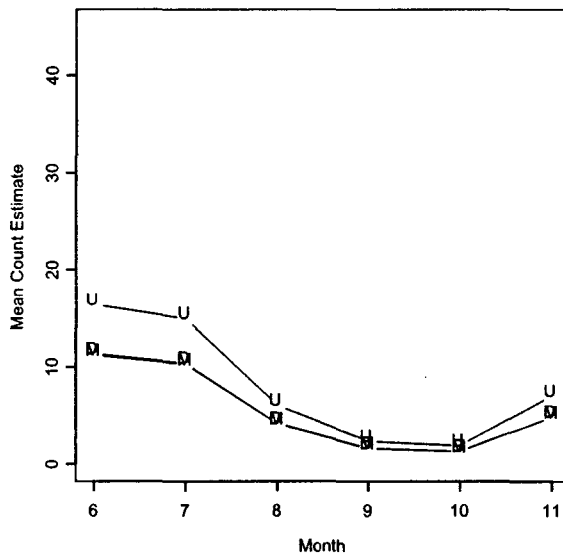


그림 3.2: GLIMMIX에 의한 점박이응애 평균밀도 추정치: U=상층부, M=중층부, D=하층부



표3.3과 표3.4는 GLIMMIX 에 의한 분석결과를 나타낸 것이다. GEE 방법과 마찬가지로 GLIMMIX 에서는 상층부의 밀도가 중층부나 하층부의 밀도보다 약  $\exp(0.38) = 1.46$  배 높은 것으로 판명된다. 그러나, 중층부와 하층부 사이의 밀도차이는 표3.4에 나타나듯이 거의 없다고 본다. 이들의 추정치들을 사용하여 시간에 따른 평균밀도의 변화를 그린 것이 그림3.2이다. 그림3.1과 비교할 때, 시간에 따른 변화추이는 서로 비슷하되, 평균밀도는 GEE 에 의한 값이 약간 높게 나타나고 있다.

PA 방법과 SS 방법을 좀 더 자세히 살펴보면, SS 의 경우, 임의효과  $u$ 가 주어졌을 때의 조건부 확률분포의 기대값은

$$E(y|u) = g^{-1}(x^T\beta + z^T u) = \exp(x^T\beta + z^T u) \tag{3.2}$$

의 모형식으로부터 출발하기 때문에,  $y$ 의 주변분포의 기대값은

$$E(y) = \exp(x^T\beta) \exp(\sigma_u^2/2) \tag{3.3}$$

가 된다(McCullach 와 Searle, 2001).

그러나, PA 방법에서는 이 값을

$$E(y) = \exp(x^T\beta_*) \tag{3.4}$$

라는 모형식을 가정하기 때문에, 만약  $\sigma_u^2 \neq 0$ 이라면, PA의  $\beta_*$  의 절편에 해당되는 모수는 SS 의  $\beta$  의 절편값보다 상수( $\sigma_u^2/2$ )만큼 증가하게 된다. 따라서, 만일 임의효과가 정규분포임을 가정하고 로그링클을 사용할 경우에, GEE 와 GLIMMIX 의 차이는 곧, 상수항의 차이로 설명되어진다. 이와 같은 현상은 두 그림, 그림3.1과 그림3.2을 비교함으로써 확인할 수 있다.

#### 4. 결론

온실장미에 서식하는 점박이용애의 분포조사를 개체특정적 접근과 모집단평균적 접근 방법으로 각각 시도해 보았다. GEE 방법이 일반화 선형모형의 구조안에서 반복측정자료를 다루는 대표적인 방법을 제공하고 있음에도 불구하고, 내포인자(nested factor)나 분할구설계(split-plot)와 같은 복잡한 형태의 혼합모형을 모형식에 표현할 수 없다는 제한적 요소를 지니고 있다. 특히, 본 연구에서는 반복이 이루어 지는 개체들이 독립이라는 GEE 의 가정은 현실성이 없어 보이기에, GEE 의 사용은 문제점을 지니고 있다고 할 수 있다.

그럼에도 불구하고, GEE 에 의한 분석결과는 추정치와 유의확률에 차이가 있을지언정 GLIMMIX 의 분석결과와 일치하고 있다. 분석결과에서 알 수 있듯이, 점박이용애의 밀도 분포는 식물체의 상층부에 높게 분포하고 있으며, 6-7월에는 높은 밀도를 가지고 있다가 8-10월에 이르러 밀도는 낮아지고, 11월 부터는 다시 증가하는 추세를 보여주고 있다.

반복측정자료분석에 대한 일반화 선형모형은 비전문통계인에게는 물론, 일반 통계분석가에게도 익숙하지 못한 분야이다. 이는 일반화 선형모형과 반복측정자료가 가지고 있는 이론적 복잡성때문이기도 하겠지만, 국내 학술지에 이들을 실제자료에 적용한 사례연구가

활발히 이루어 지지 않았기 때문이라고 생각한다. 이러한 관점에서, 본 논문은 점박이응애의 밀도분포를 추정함에 있어서 반복측정자료를 통하여 PA, SS 방법을 각각 사용, 비교하였다는 점을 강조하고 싶다. 이 논문의 분석결과 및 사용방법은 일반화선형모형을 통한 반복측정자료들에 대한 좋은 예제로 활용성이 높다고 생각한다.

## 부록

```
data a;
  input date $ plant position $ tssm time;
cards;
97-6-20 1 up 0 1
97-6-20 1 up 0 2
97-6-20 1 up 22 3
97-6-20 1 mid 6 4
. . .
proc genmod data=a;
  class plant position ;
  model tssm=position time time2 time3 /dist=poisson;
  repeated sub=plant*position /type=cs corrw;
  contrast 'pos mid to up' position 0 1 -1
  contrast 'pos down to up' position 1 0 -1
  contrast 'pos down to mid' position 1 -1 0
run;

%glimmix(data=a,
  procopt=method=reml,
  stmts=%str(
    class plant position ;
    model tssm = position time time2 time3/solution htype=1;
    random plant;
    repeated /type=toep(2) sub=plant*position rcorr;
    estimate 'pos mid to up' position 0 1 -1;
    estimate 'pos down to up' position 1 0 -1;
    estimate 'pos down to mid' position 1 -1 0;
  ), error=poisson, link=log )
run;
```

## 참고문헌

- Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, 88, 9-25.

- Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13-22.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd ed., London: Chapman and Hall.
- SAS Institute Inc. (1992). SAS technical report P-229. SAS/STAT software, proc mixed procedure, Cary, NC.
- McCullagh, P. and Searle, S. (2001). *Generalized, Linear, and Mixed Models*, John Wiley & Sons, 223-224.
- Wolfinger, R. and O'connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach", *Journal of Statistical Computation and Simulation*, **48**, 233-243.
- Zeger, S.L., Liang, K.Y. and Albert, P. (1988). Models for longitudinal data: a generalized estimating equation approach", *Biometrics*, **44**, 1049-1060

[ 2003년 8월 접수, 2003년 12월 채택 ]

## A Study on Spatial and Temporal Distribution of a Pest via Generalized Linear Mixed Models

Heungsun Park<sup>1)</sup> Kijong Cho<sup>2)</sup>

### ABSTRACT

It is an important research area in Integrated Pest Management System to estimate the pest density within plants, because the artificial controls such as spraying pesticides or biological enemies depend on the information of pest density. This paper studies the population density distribution of two-spotted spider mite in glasshouse roses. As the data were collected repeatedly on the same subject, Subject-Specific and Population Averaged approaches are used and compared.

*Keywords:* GEE, GLIMMIX, Longitudinal Data, GLMM

---

1) Professor, Department of Statistics, Hankuk University of Foreign Studies, San 79, Wangsan, Mohyun, Yongin, Kyonggi, 449-791, Korea

E-mail:hspark@hufs.ac.kr

2) Associate Professor, Division of Environmental Science and Ecological Engineering, Korea University, Seoul, 136-701

E-mail:kjcho@korea.ac.kr