YYÜ TAR BİL DERG
(YYU J AGR SCI)
2017, 27(1): 109-117

Geliş Tarihi (Received): 13.01.2017
Kabul Tarihi (Accepted): 07.03.2017

Araştırma Makalesi/*Research Article (Original Paper)*

Sayıma Dayalı Elde Edilen Verilerin Modellenmesinde Sıfır Değer Ağırlıklı Genelleştirilmiş Poisson Regresyonun Kullanılması

Süleyman SOYGÜDER, Abdullah YEŞİLOVA*, Yıldız BORA

Yüzüncü Yıl Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, Van, Türkiye
*e-posta: yesilova@yyu.edu.tr; Tel: +90 (432) 444 50 65 / 22641

Özet: Bu çalışmada, sayıma dayalı olarak elde edilen akar sayımlarının modellenmesinde sıfır değer ağırlıklı genelleştirilmiş Poisson regresyonunun uygulaması yapılmıştır. Sıfır değer ağırlıklı genelleştirilmiş Poisson regresyonunda; ortalama, aşırı yayılım ve sıfır değer ağırlıklı yayılım olmak üzere üç parametre söz konusudur. Çalışmada, aşırı yayılım ve sıfır değer ağırlıklı yayılım oldukça geniş bir aralıkta değişmiştir. Bununla birlikte aşırı yayılım ve sıfır değer ağırlıklı yayılımın akar sayımı üzerinde önemli bir etkiye sahip oldukları saptanmıştır ($p < 0.01$). Akar sayımlarının %36'sı (130 gözlem) sıfır gözlemlerden oluşmaktadır. Çalışmaya dahil edilen tüm bağımsız değişkenlerin akar sayımı üzerine olan etkileri istatistiksel olarak önemli bulunmuştur ($p < 0.05$). Akar sayımları bakımından bölgeler ve çeşitler arası farklılığın istatistiksel olarak önemli oldukları saptanmıştır ($p < 0.01$).

Anahtar kelimeler: Akar sayımı, Aşırı yayılım**,** Sıfır değer ağırlıklı veriler, Sıfır değer ağırlıklı Poisson regresyonu

Using Zero-Inflated Generalized Poisson Regression in Modelling of Count Data

Abstract: In this study zero-inflated generalized Poisson regression was applied to the modelling of mite numbers data based on count. The subjects of the zero-inflated generalized Poisson regression are three parameters as mean, overdispersion and zero-inflated dispersion. The overdispersion and zero-inflated dispersion levels range was obtained to be quite high. However, it was found that zero-inflated data and overdispersion had an important effect on mite counts ($p < 0.01$). It was obtained that 36% (130 observations) of the total numbers of mite had zero values. The effects of all independent variables were found to be statistically significant on mite counts ($p < 0.05$). The results showed that the differences among regions and varieties regarding the mite counts were statistically significant ($p < 0.01$).

Keywords: Mite counts, Overdispersion, Zero-inflated data, Zero-inflated Poisson regression

Giriş

Sayıma dayalı olarak elde edilen bağımlı değişkenin modellenmesinde Poisson regresyonu yaygın olarak kullanılmaktadır. Poisson dağılımında ortalama ile varyans birbirine eşittir. Ancak, uygulamada bu eşitliği sağlamak her zaman mümkün değildir. Bu gibi durumlarda yaygın olarak aşırı yayılım (varyansın ortalamadan büyük çıkması) ile karşılaşılmaktadır (Böhning 1998). Aşırı yayılımdan kaynaklanan etkiyi ortadan kaldırmak için genellikle negatif binomial regresyon kullanılmaktadır.

Sayıma dayalı olarak elde edilen verilerde, aşırı yayılımın yanı sıra sıfır değer ağırlıklı durumu da söz konusu olabilir. Yani, verilerin büyük bir kısmı sıfır değerlerinden oluşabilir. Sıfır değerli gözlemlerin çok sayıda olduğu sayıma dayalı olarak elde edilen verilerin modellenmesinde sıfır değer ağırlıklı regresyon yöntemleri (sıfır değer ağırlıklı Poisson regresyonu, sıfır değer ağırlıklı negatif binomial regresyonu ve Hurdle regresyonu) kullanılmaktadır.

Fazla sayıda sıfırlardan kaynaklanan etkiyi modellemek için sıfır değer ağırlıklı genelleştirilmiş Poisson (Zero-Inflated Generalized Poisson Regression = ZIGP) regresyonu son zamanlarda yaygın olarak kullanılmaktadır (Consul ve Famoye 1992; Czado ve ark. 2007; Famoye ve Singh 2003; Famoye ve Karan 2006). ZIGP kullanılarak ortalama (mean), aşırı yayılım (overdispersion) ve sıfır değer ağırlıklı (zero-

inflated) düzeyleri için ayrı ayrı regresyonlar yapılmaktadır. Böylece ortalama, aşırı yayılım ve sıfır değer ağırlıklı etkileri birbirlerinden ayrı olarak değerlendirmek mümkün olmaktadır.

Bu çalışmada amaç, sıfır değerlerinin çok olduğu ve sayıma dayalı olarak elde edilen akar sayımlarının (bağımlı değişken) modellenmesinde sıfır değer ağırlıklı genelleştirilmiş Poisson regresyonunu uygulamaktır. Bununla birlikte çalışmada kullanılacak olan ZIGP regresyon yöntemi ile diğer sıfır değer ağırlıklı regresyon yöntemlerinin performanslarının karşılaştırılması da yapılmıştır.

Materyal ve Yöntem

Materyal

Van ili Bardakçı, Şamranaltı ve Edremit bölgelerinde yaygın olarak bulunan ve Starking Delicious ve Golden Delicious elma çeşitleri üzerinde zararlı olan *Aculus schlechtendali* (Nal.)'nin populasyon yoğunluğu çalışmaları 2010–2011 yıllarında yürütülmüştür. Bu amaçla, her ağaç için 25, her bölge için toplam 125 adet yaprak ve her yaprağın 3cm²lik alanında sayım yapılmıştır. Farklı zamanlardaki *M. communis* L. üzerinde gözlenen *A. schlechtendali*'nin protogyne, deutogyne dönemleri ile yumurta ve nymphopupa dönemleri belirlenmeye çalışılmıştır. Toplanan yapraklar üzerinde *Zetzellia mali* (Ewing)'nin populasyon sayımı yapılmıştır. Ayrıca, söz konusu aylara ilişkin sıcaklık ve nem değerleri de çalışmaya dahil edilmiştir.

Yöntem

Sıfır Değer Ağırlıklı Genelleştirilmiş Poisson Regresyonu

Y bağımlı değişkeni μ, φ ve ω parametreleri ile birlikte sıfır değer ağırlıklı genelleştirilmiş Poisson dağılımı gösterir ve $Y: \text{ZIGP}(\mu,\phi,\omega)$ şeklinde gösterilmektedir. Burada μ, φ ve ω parametreleri sırasıyla ortalama, aşırı yayılım ve sıfır değer ağırlıklı parametrelerini göstermektedir. Sıfır değer ağırlıklı genelleştirilmiş Poisson dağılımının yoğunluk fonksiyonu aşağıdaki gibi yazılabilir (Consul ve Famoye 1992; Czado ve ark. 2007),

$$P(Y=y\mid\mu,\varphi,\omega)=I_{\{y=0\}}\left[\omega+(1-\omega)e^{-\frac{\mu}{\varphi}}\right]+I_{\{y>0\}}\left[(1-\omega)\frac{\mu(\mu+(\varphi-1)y)^{y-1}}{y!}\varphi^{-y}e^{-\frac{1}{\varphi}(\mu+(\varphi-1)y)}\right] \quad (1)$$

Bazı veri setlerinde bir sabit aşırı yayılım ve/veya sıfır değer ağırlıklı parametreler sınırlanabilir. Böyle durumlarda, aşırı yayılım ya da sıfır değer ağırlıklı parametrelerinden hangilerinin değişip değişmediğini sıfır değer ağırlıklı genelleştirilmiş Poisson regresyonu kullanılarak belirlenebilir. ZIGP regresyon modelinde $X_i=(1,x_{i1},x_{i2},....x_{ip})^t$, $\varphi_i=(1,\varphi_{i1},\varphi_{i2},....\varphi_{ir})^t$ ve $Z_i=(1,z_{i1},z_{i2},....z_{iq})^t$ vektörleri sırasıyla; ortalama, aşırı yayılım ve sıfır yayılım için bağımsız değişkenleri göstersin. $\text{ZIGP}(\mu_i,\varphi_i,\omega_i)$ modeli şansa bağlı, sistematik ve parametrik olmak üzere üç bileşenden oluşmaktadır (Czado ve ark. 2007). Söz konusu bu bileşenler veri setimize uyarlanarak aşağıdaki gibi yazılabilir,

Şansa bağlı bileşen

$\{Y_i, 1\le i\le n\}$ bağımlı değişken olan akar sayımları birbirinden bağımsız ve $Y_i \sim \text{ZIGP}(\mu_i,\varphi_i,\omega_i)$ dağılımı göstermektedir.

Sistematik bileşen

Burada, bağımlı değişken ile bağımsız değişkenler arasında doğrusal ilişkiyi sağlamak için bağlantı (link) fonksiyonu kullanılmaktadır. Ortalama, aşırı yayılım ve sıfır değer ağırlıklı yayılım için, Y_i bağımlı değişkeni üzerinde $\eta_i^{\mu}(\beta) = x_i^t\beta$, $\eta_i^{\varphi}(\alpha) = \omega_i^t\alpha$, $\eta_i^{\omega}(\gamma) = z_i^t\gamma$ gibi üç doğrusal tahmin edici etkili olmaktadırlar. Burada $\beta = (\beta_0, \beta_1, ..., \beta_p)^t$, $\alpha = (\alpha_0, \alpha_1, ..., \alpha_r)^t$, $\gamma = (\gamma_0, \gamma_1, ..., \gamma_q)^t$ bilinmeyen regresyon parametreleri, $X_i = (x_1, x_2, ..., x_n)^t$, $W_i = (w_1, w_2, ..., w_n)^t$, $Z_i = (z_1, z_2, ..., z_n)^t$ desen matrisleri olarak adlandırılmaktadır.

Parametrik bağlantı bileşeni

$\eta_i^{\mu}(\beta)$, $\eta_i^{\varphi}(\alpha)$, $\eta_i^{\omega}(\gamma)$ doğrusal tahmin edicileri ile $\mu_i(\beta)$, $\varphi_i(\alpha)$, $\omega_i(\gamma)$ (i=1,…,n) parametreleri arasındaki doğrusal fonksiyonlar aşağıdaki gibi verilebilir.

Ortalama düzeyi:

$$\begin{aligned} &E(Y_i \mid \beta) = \mu_i(\beta) = E_i e^{x_i^t\beta} = e^{x_i^t\beta + \log(E_i)} > 0 \\ &\quad \Leftrightarrow \eta_i^{\mu}(\beta) = \log(\mu_i(\beta)) - \log(E_i) \text{ (log bağlantı)} \end{aligned} \tag{2}$$

Aşırı yayılım düzeyi:

$\varphi_i(\alpha) = 1$ olması veri setinde aşırı yayılım olmadığı anlamına gelmektedir. Bununla birlikte, $\varphi_i(\alpha) > 1$ olması durumunda veri setinde aşırı yayılım söz konusudur. Bu aşırı yayılım aşağıdaki gibi modellenmektedir.

$$\varphi_i(\alpha) = 1 + e^{w_i^t\alpha} > 1$$
$$\eta_i^{\varphi}(\alpha) = \log(\varphi_i(\alpha) - 1) \text{ (shifted log bağlantı)}$$

Her bir bağımsız değişken için aşırı yayılım düzeyi aşağıdaki gibi hesaplanabilir;

$$\hat{V}(X = (X = x, W = w, Z = z) = \varphi^2\mu.\omega = \hat{\varphi}(w)^2 + \hat{\mu}(x)\hat{\omega}(z) \tag{3}$$

Sıfır değer ağırlıklı düzey:

$$\omega_i(\gamma) = \frac{e^{z_i^t\gamma}}{1 + e^{z_i^t\gamma}} \in (0,1)$$
$$\Leftrightarrow \eta_i^{\omega}(\gamma) = \log\left(\frac{\omega_i(\gamma)}{1 - \omega_i(\gamma)}\right) \text{ (logit bağlantı)}$$

Her bir bağımsız değişken için sıfır değer ağırlıklı düzey aşağıdaki gibi hesaplanabilir;

$$\begin{aligned} &\hat{P}(Y = 0 / (X = x, W = w, Z = z) = \hat{\omega}(z) + (1 - \hat{\omega}(z))\exp\left(\frac{-\hat{\mu}(x)}{\hat{\varphi}(w)}\right) \\ &\hat{\omega}(z) = \frac{\exp(\hat{\gamma}_0 + z_1\hat{\gamma}_1 + ... + z_q\hat{\gamma}_q)}{1 + \exp(\hat{\gamma}_0 + z_1\hat{\gamma}_1 + ... + z_q\hat{\gamma}_q)} \end{aligned} \tag{4}$$

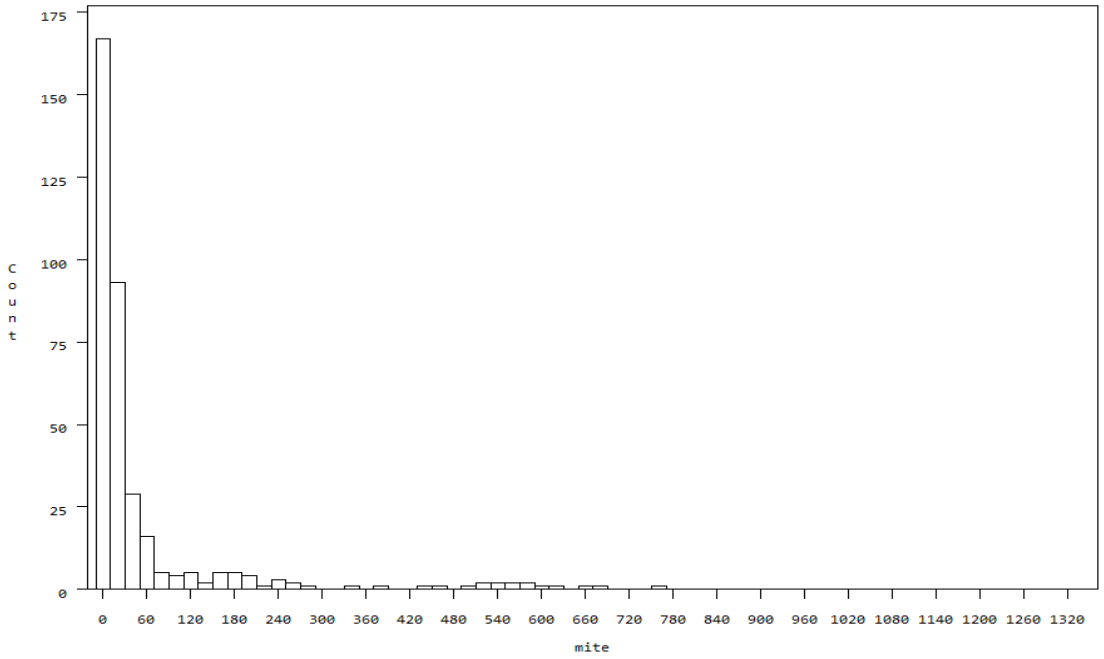
$(\beta^t, \alpha^t, \gamma^t)$ bilinmeyen parametreler göstermektedir. y_i bağımlı değişkeni için ZIGP regresyonunun log olabilirlik fonksiyonu[7],

$$
\begin{aligned}
l(\delta) = \sum_{i=1}^{n} I_{(y_i=0)} & \left[\log\left(e^{z_i^t\gamma} + \exp\left(-\frac{E_i . e^{x_i^t\beta}}{1+e^{\omega_i^t\alpha}} \right) \right) - \log(1+e^{z_i^t\gamma}) \right] \\
& + I_{(y_i>0)} \left[-\log(1+e^{z_i^t\gamma}) + \log(E_i) + x_i^t\beta - \log(y_i!) - y_i \log(1+e^{\omega_i^t\alpha}) \right. \\
& \left. + (y_i - 1)\log(E_i e^{x_i^t\beta} + e^{\omega_i^t\alpha} y_i) - \frac{E_i e^{x_i^t\beta} + e^{\omega_i^t\alpha} y_i}{1+e^{\omega_i^t\alpha}} \right]
\end{aligned}
\tag{5}
$$

biçiminde yazılabilir. Eşitlik 5'te ortalama, aşırı yayılım ve sıfır değer ağırlıklı yayılım $(\beta^t, \alpha^t, \gamma^t)$ bilinmeyen parametreleri olabilirlik fonksiyonu maksimize edilerek en çok olabilirlik yöntemi kullanılarak elde edilmektedirler (Czado ve ark. 2007).

Bulgular ve Tartışma

Bu çalışmada, gerekli istatistiksel analizler R 2.10.1 istatistik yazılım programı kullanılarak yapılmıştır. Bölgeler, yıllar, aylar, çeşitler, sıcaklık ve nem modele bağımsız değişken olarak, akar sayımları ise bağımlı değişken olarak dahil edilmiştir. Akar sayımlarının grafiği Şekil 1'de verilmiştir. Akar sayımlarının grafiği aşırı yayılım ve sıfır değer ağırlıklı gözlemlerden dolayı oldukça sağa doğru çarpık olmuştur. Çalışmada, gözlem değerlerinin %36'sı (130 gözlem) sıfır değerli olmuştur. Bu tür veriler dönüşümlere tabi tutulmalarına rağmen sağa doğru aşırı çarpıklık çok fazla değişmemektedir. Burada, akar sayıları karekök dönüşümüne tabi tutulduktan sonra bile dağılışın şekli yine sağa çarpık olmuştur. Bu durum parametrik testlerdeki Normal dağılım varsayımını karşılamamaktadır. Buna rağmen Poisson dağılışı gösteren bu tip verilere doğrusal modelleri uygulamak, doğru olmayan parametre tahminleri ve standart hataların elde edilmesine neden olabilir (McCullagh ve Nelder 1989).



Şekil 1. Akar Sayımlarının Grafiği

Yukarıda da ifade edildiği gibi bağımlı değişkenin Poisson dağılışı göstermesi durumunda Poisson regresyonunun uygulanması gerekmektedir. Ancak bağımlı değişkende (akar sayımları), eğer gerçekten bir

aşırı yayılım durumu varsa Poisson regresyonu yerine, söz konusu aşırı yayılımı modelleyen negatif binomial regresyonunun kullanılması gerekmektedir (Cox 1983; Ridout ve ark. 2001; McCullagh ve Nelder 1989).

Poisson (PR), negative binomial (NBR), sıfır değer ağırlıklı (ZIP), sıfır değer ağırlıklı negative binomial (ZINB), Poisson Hurdle (PH), negative binomial Hurdle (NBH), genelleştirilmiş Poisson (GP) ve ZIGP regresyonları için Akaiki bilgi ölçütü (AIC) değerleri çizelge 1'de verilmiştir. $PR(\mu_i)$ regresyon modelinden elde edilen AIC uyum ölçütü diğer regresyon modellerine göre oldukça yüksek bulunmuştur. Bunun nedeninin veri setindeki aşırı yayılım ve sıfır değerli gözlemlerin olduğu söylenebilir. Çizelge 1'de veri setini en iyi açıklayan regresyon modelinin $ZIGP(\mu_i,\varphi_i,\omega_i)$ olduğu saptanmıştır. Çünkü $ZIGP(\mu_i,\varphi_i,\omega_i)$ modelinden elde edilen AIC değeri diğer modellere göre en küçük olarak elde edilmiştir. Bu regresyon modeline ek olarak $ZIGP(\mu_i,\varphi_i,\omega)$ ve $ZIGP(\mu_i,\varphi,\omega_i)$ modellerine ilişkin AIC değerleri de Çizelge 1'de verilmiştir. $ZIGP(\mu_i,\varphi,\omega)$ modeli ile $ZIGP(\mu_i,\varphi,\omega_i)$ arasındaki fark, sıfır değer ağırlıklı parametrenin (ω) değişkenlik göstermesidir. Benzer şekilde, $ZIGP(\mu_i,\varphi,\omega)$ modeli ile $ZIGP(\mu_i,\varphi_i,\omega)$ arasındaki fark, aşırı yayılım parametresinin (φ) değişkenlik göstermesidir. Çizelge 1 kullanılarak aşırı yayılım ve sıfır değer ağırlıklı yayılım için model karşılaştırılması yapılabilir.

İlk olarak, sıfır değer ağırlıklı parametrenin önemli olup olmadığını belirlemek için iç içe modeller karşılaştırılmıştır. Bu nedenle, $PR(\mu_i)$ile $ZIP(\mu_i,\omega)$, $NBR(\mu_i)$ ile $ZINB(\mu_i,\omega)$ ve GP ile $ZIGP(\mu_i,\varphi,\omega)$ regresyon modelleri karşılaştırılmıştır. $PR(\mu_i)$ile $ZIP(\mu_i,\omega)$ karşılaştırıldığında, AIC değeri 9328'den 3997'ye düşmüştür. Buna paralel olarak Voung istatistiği 8.33 olarak elde edilmiştir. Böylece sıfır değer ağırlıklı parametreye göre; $ZIP(\mu_i,\omega)$ modeli $PR(\mu_i)$ modeline tercih edilmiştir ($p<0.01$). $NBR(\mu_i)$ ve $ZINB(\mu_i,\omega)$ modelleri karşılaştırıldığında, AIC değeri 4043'ten 1418'e düşmüş ve Voung istatistiği 5.91 olarak elde edilmiştir. Böylece; hem AIC hem de Voung istatistiği sonucu, sıfır değer ağırlıklı parametreye göre $ZINB(\mu_i,\omega)$ modeli $NBR(\mu_i)$ modeline tercih edilmiştir ($p<0.01$). Son olarak, GP ile $ZIGP(\mu_i,\varphi,\omega)$ regresyon modelleri karşılaştırıldığında, hem AIC hem de Voung istatistiği $ZIGP(\mu_i,\varphi,\omega)$ modelinin GP modeline tercih edilmesi gerektiğini göstermektedir ($p<0.01$). Burada sıfır değer ağırlıklı parametre için yapılan 3 ayrı model karşılaştırması sonucunda, veri setindeki sıfır değer ağırlıklı yayılımın istatistiksel olarak önemli olduğunu göstermektedir.

Yukarıda, sıfır değer ağırlıklı parametre (ω) için yapılan model karşılaştırmaları önemli bulunmuştur. Şimdi ise, sabit sıfır değer ağırlıklı yayılım (ω) ile değişkenlik gösteren sıfır değer ağırlıklı yayılım (ω_i) parametrelerini karşılaştırmak gerekmektedir. Böylece sıfır değer ağırlıklı parametrenin değişkenlik gösterip göstermediği saptanabilir. Bu nedenle, iç içe modeller olan $ZIP(\mu_i,\omega)$ile $ZIP(\mu_i,\omega_i)$, $ZINB(\mu_i,\omega)$ile $ZINB(\mu_i,\omega_i)$ ve $ZIGP(\mu_i,\varphi,\omega)$ ile $ZIGP(\mu_i,\varphi,\omega_i)$ modelleri karşılaştırılmıştır. $ZIP(\mu_i,\omega)$ve $ZIP(\mu_i,\omega_i)$modelleri karşılaştırıldığında, AIC değeri 3997'den 1042'ye düşmüştür. Buna paralel olarak Voung istatistiği 5.49 olarak elde edilmiştir. Böylece değişkenlik gösteren sıfır değer ağırlıklı parametreye göre; $ZIP(\mu_i,\omega_i)$ modeli $ZIP(\mu_i,\omega)$ modeline tercih edilmiştir ($p<0.01$). $ZINB(\mu_i,\omega)$ ile $ZINB(\mu_i,\omega_i)$ regresyon modelleri karşılaştırıldığında, AIC değeri 1418'den 1010'a düşmüş ve buna paralel olarak Voung istatistiği 3.47 olarak elde edilmiştir. Böylece değişkenlik gösteren sıfır değer ağırlıklı parametreye göre; $ZINB(\mu_i,\omega_i)$modeli $ZINB(\mu_i,\omega)$ modeline tercih edilmiştir ($p<0.05$). Son olarak, $ZIGP(\mu_i,\varphi,\omega)$ ile $ZIGP(\mu_i,\varphi,\omega_i)$ modelleri karşılaştırılmıştır. Hem AIC hem de Voung istatistiği $ZIGP(\mu_i,\varphi,\omega_i)$ modelinin $ZIGP(\mu_i,\varphi,\omega)$modeline tercih edilmesi gerektiğini göstermektedir ($p<0.05$). Yapılan 3 farklı model karşılaştırması sonucunda, sıfır değer ağırlıklı yayılım parametresinin sabit olmayıp değişkenlik gösterdiği saptanmıştır.

Aşırı yayılım parametresinin(φ) önemli olup olmadığını belirlemek için PR modeli ile NBR modelini karşılaştırdık. Elde edilen sonuçlara göre, AIC değeri 9328'den 4043'e düşmüştür. Buna paralel olarak Voung istatistiği 8.07 olarak elde edilmiştir. Her iki istatistik değeri, veri setindeki aşırı yayılım parametresinin istatistiksel olarak önemli olduğunu göstermiştir ($p < 0.01$). Aşırı yayılımın etkisi önemli çıktıktan sonra, sabit etkili aşırı yayılım (φ) ile değişen aşırı yayılım (φ_i) parametrelerini karşılaştırdık. Bunun için PR ile GP ve $\mathbf{ZIGP(\mu_i,\varphi,\omega)}$ ile $ZIGP(\mu_i,\varphi_i,\omega)$ modelleri esas alındığında; PR ile GP modelleri karşılaştırıldığında, AIC değeri 9328'den 2681'e düşerken, Voung istatistiği 7.63 olarak elde edilmiştir. Bu nedenle GP modeli PR modeline tercih edilmiştir. Benzer şekilde $\mathbf{ZIGP(\mu_i,\varphi,\omega)}$ ile $ZIGP(\mu_i,\varphi_i,\omega)$ modelleri karşılaştırıldığında hem AIC hem de Voung istatistikleri, $ZIGP(\mu_i,\varphi_i,\omega)$ modelinin $\mathbf{ZIGP(\mu_i,\varphi,\omega)}$ modelinden daha iyi sonuç verdiği saptanmıştır. Bu bulgulara göre değişen aşırı yayılımın (φ_i) istatistiksel olarak önemli olduğu saptanmıştır.

Yapılan tüm model karşılaştırmalarından sonra, $ZIGP(\mu_i,\varphi_i,\omega_i)$ modelinin diğer tüm regresyon modellerinden daha iyi sonuç veridiği saptanmıştır (Çizelge 1). Bu nedenle parametre tahminleri $ZIGP(\mu_i,\varphi_i,\omega_i)$ regresyon modeli esas alınarak yorumlanmıştır. $ZIGP(\mu_i,\varphi_i,\omega_i)$ için parametre tahminleri çizelge 2'de verilmiştir.

Çizelge 1. Farklı regresyon modelleri için AIC değerleri

Model	**AIC**
Poisson regression($\mathbf{PR(\mu_i)}$)	9328
Negative binomial regresyonu $NBR(\mu_i)$	4043
Zero-inflated Poisson regresyonu $\mathbf{ZIP(\mu_i,\omega)}$	3997
Zero-inflated Poisson regresyonu $ZIP(\mu_i,\omega_i)$	1042
Zero-inflated negative binomial regresyonu $ZINB(\mu_i,\omega)$	1418
Zero-inflated negative binomial regresyonu $ZINB(\mu_i,\omega_i)$	1010
Poisson Hurdle regresyonu $PH(\mu_i,\omega)$	2857
Negative binomial Hurdle regresyonu $NBH(\mu_i,\omega)$	1057
Generalized Poisson regresyonu $\mathbf{GP(\mu_i,\varphi)}$	2681
Zero-inflated generalized Poisson regresyonu $\mathbf{ZIGP(\mu_i,\varphi,\omega)}$	1069
Zero-inflated generalized Poisson regresyonu $ZIGP(\mu_i,\varphi_i,\omega)$	1002
Zero-inflated generalized Poisson regresyonu $ZIGP(\mu_i,\varphi,\omega_i)$	1010
Zero-inflated generalized Poisson regresyonu $ZIGP(\mu_i,\varphi_i,\omega_i)$	974

Aşırı yayılım ve sıfır değer ağırlıklı gözlemlerden dolayı, Çizelge 2'de verilen parametre tahminleri farklı bulunmuştur. Yani, bağımsız değişkenler için elde edilen tahmin değerleri ortalama regresyon, aşırı yayılım regresyonu ve sıfır değer ağırlıklı regresyonlarda oldukça farklı çıkmışlardır. Bu durumda aşırı yayılımın ve sıfır değer ağırlıklı gözlemlerin, parametre tahminleri üzerinde ne kadar etkili olduğunu göstermiştir.

Çizelge 2'de verilen ortalama regresyon için parametre tahminlerine bakıldığında; akar sayımları üzerine tüm bağımsız değişkenlerin etkisi istatistiksel olarak önemli bulunmuştur ($p < 0.01$). Çizelge 2'de ortalama regresyon modeli için verilen parametre tahminlerinin yorumlanması doğrusal regresyona göre farklıdır. Bunun için aşağıda verilen eşitlik 6 kullanılarak her bir bağımsız değişkenin akar sayımları üzerine olan etkisi yorumlanabilir. Yani aşağıda verilen eşitlikteki log dönüşüm kullanılarak parametre tahminleri yorumlanmaktadır. Örneğin, bölgeler arası farklılık, akar sayımlarında 0.264 (%26.4)'lük bir azalışa neden olmuştur. Buna göre akar sayımları bölgelere göre farklılık göstermiştir ($p < 0.01$). Sıcaklığın bir birim artması akar sayımlarında 0.802 (%80.2)'lik bir artışa neden olmuştur ($p < 0.01$). Çeşitler arası farklılık akar

sayımlarında 0.095 (%9.5)'lik bir değişime neden olduğu saptanmıştır ve bu değişim istatistiksel olarak önemli bulunmuştur ($p < 0.05$). Aylar arası farklılık akar sayımlarında 0.192 (%19.2)'lik bir azalmaya neden olmuştur ($p < 0.01$).

$$\log(akar \quad sayımı) = 828.051 + 0.411 * \text{yıl} - 0.307 * \text{bölge} + 0.589 * \text{sıcaklık} + 0.133 * \text{nem} - 0.213 * \text{ay} - 0.091 * \text{çeşit} \quad (6)$$

Sıfır değer ağılıklı regresyonda; yıl değişkeni hariç diğer tüm bağımsız değişkenlerin akar sayımları üzerine olan etkileri istatistiksel olarak önemli bulunmuştur ($p < 0.05$). Aşırı yayılım regresyonunda; akar sayımları üzerine sıcaklık, nem ve bölgelerin etkisi önemli bulunmuşken ($p < 0.01$), yılların, ayların ve çeşitlerin etkileri önemsiz bulunmuşlardır ($p > 0.05$).

Çizelge 2. $ZIGP(\mu_i, \varphi_i, \omega_i)$ regresyon modeli için parametre tahminleri

	Ortalama regresyon (Mean regression)		**Sıfır değer ağırlıklı regresyon (Zero-inflated regression)**		**Aşırı yayılım regresyon (Overdispersion regression)**	
Bağımsız değişkenler	Parametre tahmini	Standard hata	Parametre tahmini	Standard hata	Parametre tahmini	Standard hata
Sabit	828.051	81.118**	1022.817	158.509**	273.911	34.608**
Yıl	0.411	0.043**	-0.241	0.073	0.094	0.006
Bölge	-0.307	0.018**	0.406	0.008*	-0.563	0.104*
Sıcaklık	0.589	0.061**	0.753	0.081**	0.328	0.075*
Nem	0.133	0.013**	0.102	0.028**	0.216	0.043*
Ay	0.213	0.042**	0.463	0.012*	-0.471	0.028
Çeşit	0.091	0.007*	0.321	0.091*	-0.152	0.072
Ortalamanın değişim aralığı **($\hat{\mu}$)**					(7.830, 191.130)	
Sıfır değer ağırlıklı parametrenin değişim aralığı **($\hat{\omega}$)**					(0, 0. 41)	
Aşırı yayılım parametresinin değişim aralığı **($\hat{\omega}$)**					(8.416, 213.047)	

*p<0.05 **p<0.01

Aylar ve bölgeler gibi kategorik bağımsız değişkenlerin her bir düzeyi için sıfır değer ağırlıklı yayılım ve aşırı yayılım miktarları, yöntem bölümünde verilen eşitlik 3 ve 4 kullanılarak aşağıdaki gibi hesaplanmıştır. Elde edilen sonuçlar çizelge 3 ve çizelge 4'te verilmiştir.

Çizelge 3'te, aylar için aşırı yayılım düzeyi ($\hat{V}(X, W, Z)$) 1.700 ile 34.340 arasında değişmiştir. 2010 yılı için en yüksek aşırı yayılım değerleri sırasıyla Temmuz ve Ağustos aylarında elde edilmiştir. Bunun nedeninin bu aylardaki yüksek sıcaklık ve nemden kaynaklandığı söylenebilir. En düşük aşırı yayılım değeri Haziran ve Eylül aylarında görülmekle birlikte, Mayıs ve Ekim aylarında hiç aşırı yayılım saptanmamıştır. Genellikle sıcaklığın düşük olduğu bu aylarda hiç akar sayımlarına rastlanamamıştır. Bu iki ayda çoğunlukla sıfır değerli veriler gözlenmiştir. 2010 yılı için sıfır değer ağırlıklı yayılım Mayıs ve Ekim aylarında maksimum; Haziran, Temmuz ve Ağustos aylarında ise minimum olarak gözlenmiştir.

Çizelge 3'te 2011 yılı esas alındığında, aylar için aşırı yayılım düzeyi ($\hat{V}(X, W, Z)$) 1.738 ile 58.277 arasında değişmiştir. 2011 yılı için en yüksek aşırı yayılım değeri Temmuz ayında elde edilmiştir. En düşük aşırı yayılım değeri Haziran, Ağustos ve Eylül aylarında görülmüştür. Mayıs ve Ekim aylarında hiç aşırı yayılım saptanmamıştır. Genellikle sıcaklığın düşük olduğu bu aylarda hiç akar sayımlarına rastlanamamıştır. Bu iki ayda çoğunlukla sıfır değerli veriler gözlenmiştir. 2011 yılı için sıfır değer ağırlıklı yayılım Mayıs ve Ekim aylarında maksimum; Haziran, Temmuz ve Ağustos aylarında ise minimum olarak gözlenmiştir. Eylül ayındaki sıfır değer ağırlıklı yayılım %2.139 olarak tahmin edilmiştir.

Çizelge 4'te 2010 yılı esas alındığında, aylar için aşırı yayılım düzeyi ($\hat{V}(X, W, Z)$) 2.620 ile 5.853 arasında değişmiştir. 2010 yılı için en yüksek aşırı yayılım değeri Bardakçı'da elde edilmiştir. En düşük aşırı yayılım değeri ise Edremit'te görülmüştür. 2010 yılı için aşırı yayılım düzeyi Edremit ve Şamranaltı'nda benzerlik göstermiştir. 2011 yılı için bölgelere göre tahmin edilen aşırı yayılım düzeyi bakımından Edremit

bölgesinin daha yüksek olduğu belirlenmiştir. 2010 yılında tahmin edilen sıfır değer ağırlıklı düzeylerin birbirine yakın oldukları saptanmıştır. Bardakçı'daki sıfır değer ağırlıklı parametre değerinin diğer iki bölgeye nazaran daha yüksek olduğu söylenebilir. 2011 yılında tahmin edilen sıfır değer ağırlıklı düzeylerin birbirine yakın oldukları saptanmıştır. Edremit'te tahmin edilen sıfır değer ağırlıklı parametre değerinin diğer iki bölgeye nazaran daha yüksek olduğu söylenebilir.

Çizelge 3. Her bir ay için tahmin edilen aşırı yayılım ve sıfır değer ağırlıklı parametre tahminleri

	2010			**2011**	
Aylar	$\hat{V}(X,W,Z)$	$\hat{P}(Y=0/x,w,z)$	Aylar	$\hat{V}(X,W,Z)$	$\hat{P}(Y=0/x,w,z)$
Mayıs	-	%100	Mayıs	-	%100
Haziran	1.700	-	Haziran	3.890	-
Temmuz	29.339	-	Temmuz	58.277	-
Ağustos	34.340	-	Ağustos	3.762	-
Eylül	1.753	%1.78	Eylül	1.738	%2.139
Ekim	-	%100	Ekim	-	%100

Çizelge 4. Her bir bölge için tahmin edilen aşırı yayılım ve sıfır değer ağırlıklı parametre tahminleri

	2010			**2011**	
Bölge	$\hat{V}(X=(X,W,Z)$	$\hat{P}(Y=0/x,w,z)$	Bölge	$\hat{V}(X=(X,W,Z)$	$\hat{P}(Y=0/x,w,z)$
Bardakçı	5.853	%33.300	Bardakçı	4.094	%30.461
Şamranaltı	3.918	%29.371	Şamranaltı	4.408	%29.018
Edremit	2.620	%31.283	Edremit	6.838	%32.430

Bu çalışmada, veri setindeki aşırı yayılım ve fazla sayıdaki sıfır değerlerinden dolayı Poisson Regresyonu için AIC istatistiği diğer tüm regresyon modellerine nazaran oldukça yüksek bulunmuştur. Bununla birlikte en küçük AIC değeri, hem aşırı yayılım hem de sıfır değer ağırlıklı parametrelerin değişkenlik gösterdiği $ZIGP(\mu_i,\varphi_i,\omega_i)$ modelinde elde edilmiştir. Yapılan bu tip çalışmalarda, ZIGP regresyonun olmadığı durumlarda, genellikle en iyi sonuç veren regresyon modelinin ZINB olduğu bilinmektedir (Consul ve Famoye 1992; Famoye ve Singh 2003; Famoye ve Karan 2006; Czado ve ark. 2007; Zamani ve Ismail 2014; Zhao ve ark. 2014).

Akar sayımlarının ortalaması ve varyansı sırasıyla, 64.748 ve 742.061 olarak elde edilmiştir. Bu iki istatistik arasındaki fark, veri setinde hem sıfır değer ağırlıklı gözlemlerin hem de aşırı yayılımın ne kadar etkili olduğunun kanıtıdır. Çizelge 2'de aşırı yayılımın değişim aralığının 7.830 ile 191.130 arasında oldukça yüksek olarak tahmin edilmişti. Aşırı yayılımın en büyük nedenlerinden biri de sıcaklık ve nem gibi çevresel etkilerden kaynaklanmasıdır (Kasap 2010; Yeşilova ve ark. 2011).

Çizelge 2'de akar sayımlarının %36'sı (130 gözlem) sıfır değerli olduğu belirlenmiştir. Bununla birlikte sıfır değer ağırlıklı parametrenin değişim aralığı %0 ile %41 arasında olmuştur. Bitki koruma ile ilgili yapılan bu tip çalışmalarda, gerek akar sayımları olsun, gerekse yumurta sayımları olsun aşırı yayılım ve sıfır değer ağırlıklı gözlemler bakımından çok büyük değer alabilmektedir (Yeşilova ve ark. 2011).

Sonuç

Veri setindeki aşırı yayılım ve sıfır değerli gözlemlerin çok yüksek olmasından dolayı $PR(\mu_i)$ regresyon modelinden elde edilen AIC uyum ölçütü diğer regresyon modellerine göre oldukça yüksek elde edilmiştir. AIC ölçütüne bakıldığında en iyi regresyon modelinin $ZIGP(\mu_i,\varphi_i,\omega_i)$ olduğu saptanmıştır. Bu regresyon modelinde hem aşırı yayılım hem de sıfır değer ağırlıklı yayılım parametreleri değişkenlik göstermektedir. Bu sonuç, veri setinde aşırı yayılım ve sıfır değer ağırlıklı yayılım parametrelerinin ne kadar yüksek bir aralıkta değiştiğini göstermiştir. Bu nedenle parametre tahminleri $ZIGP(\mu_i,\varphi_i,\omega_i)$ regresyon modeli esas alınarak yapılmıştır. Bununla birlikte çok daha karmaşık bu tip veri setlerinin

analizinde, söz konusu hetrojenliği modellemek için sıfır değer ağırlıklı karışımlı Poisson regresyonu modellerinin (zero-inflated mixture Poisson regression) son zamanlarda yaygın olarak kullanılmaktadır.

Teşekkür

Bu çalışma Yüzüncü Yıl Üniversitesi Bilimsel araştırma ve Projleri Başkanlığı tarafından desteklenen 2015-FBY-YL187 nolu ve "Sıfır Gözlemlerinin Çok Fazla Sayıda Olduğu Sayıma Dayalı Olarak Elde Edilen Verilerin Modellenmesinde Sıfır Değer Ağırlıklı Geneleştirilmiş Poisson Regresyonunun Kullanılması" isimli Yüksek lisans tez projenin bir kısmıdır.

Kaynaklar

Böhning D (1998). Zero- Inflated Poisson Models and C.A.MAN: A Tutorial Collection of Evidence. Biometrical Journal, 40(7): 833-843.

Consul P, Famoye F (1992). Generalized Poisson regression model. Comn. Statist. The Met, 21(1): 89–109.

Cox R (1983). Some Remarks on Overdispersion. Biometrika, 70: 269-274.

Czado C, Erhardt V, Min A, Wagner S (2007). Dispersion and zero-inflation level applied to patent outsourcing rates Zero-inflated generalized Poisson models with regression effects on the mean. Statistical Modelling, 7(2): 125-153.

Famoye F, Singh K.P (2003). On inflated generalized Poisson regression models. Advanced Applied Statistics, 3(2): 145–158.

Famoye F, Karan P.S (2006). Zero- inflated generalized Poisson regression model with an application to domestic violence data. Journal of Data Science, 5(4): 117-130.

Kasap İ (2010). Seasonal Population Development of Spider Mites (Acari: Tetranychidae) and Their Predators in Sprayed and Unsprayed Apple Orchards in Van, Turkey. XIII International Congress of Acarology | Recife, Pernambuco, Brazil – August 23-27, 2010.

Lambert D (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1): 1-13.

Luo J, Qu Y (2013). Analysis of hypoglycemic events using negative binomial models. Pharm Stat., 12(4): 233-42.

McCullagh P, Nelder J.A (1989). Generalized Linear Models. Second Edition, London, UK, Chapmann and Hall.

Ridout M, Hinde J, Demetrio C.G.B (2001). A score test for a zero-inflated Poisson regression model against zero-inflated negative binomial alteratves. Biometrics. 57: 219-233.

Yeşilova A, Özgökçe, M.S, Atlıhan R, Karaca İ, Özgökçe F, Yıldız Ş, Kaya Y (2011). Sıfır değer ağırlıklı genelleştirilmiş Poisson regresyonu yardımıyla Van Gölü'nde Notonecta viridis Delcourt, 1909 (Hemiptera: Notonectidae)'in populasyon değişimi üzerinde fiziko-kimyasal çevresel koşulların etkilerinin araştırılması. Turkish Journal of Entomology, 35(2): 325-338.

Zamani H, Ismail N (2014). Functional form for the zero-inflated generalized Poisson regression model. Communication in Statistics-Theory and Methods, 43(3): 515-529.

Zhao W, Zhang R, Liu J, Lv Y (2014). Semi varying coefficient zero-inflated generalized Poisson regression model. Communication in Statistics-Theory and Methods, 44(1): 171-185.