Open Section

DOI: 10.25115/eea.v40i3.6927

Studies of Applied Economics / Estudios de Economía Aplicada

Volume: 40-3, September 2022 // ISSN: 1133-3197

A modelling of the number of *almazaras* by municipality in Andalusia

VALENTINA CUEVA-LÓPEZ[1], JOSÉ RODRÍGUEZ-AVI[2], MARÍA JOSÉ OLMO-JIMÉNEZ[3], JULIA RODRÍGUEZ-REINOSO[4]

[1]Departamento de Estadística e I. O., UNIVERSITY OF JAÉN. SPAIN. E-mail: vcueva@ujaen.es
[2]Departamento de Estadística e I. O., UNIVERSITY OF JAÉN. SPAIN. E-mail: jravi@ujaen.es
[3]Departamento de Estadística e I. O., UNIVERSITY OF JAÉN. SPAIN. E-mail: mjolmo@ujaen.es
[4] E-mail: juliarodriguezreinoso@gmail.com

ABSTRACT

An *almazara* (oil mill) is an essential piece in the production of olive oil since it is the place where the olive is milled and the olive oil is obtained. They are usually linked to producer cooperatives. They are structures that require specialized machinery and that on multiple occasions are underutilised, given the presence of several of them at very close distances. In addition, they characterise the mainly olive grove municipalities and their study provides a valuable information of economic interest.

From a statistical point of view, the "number of oil mills per municipality" is a count data variable that exhibits overdispersion. In this study, we focus on the oil mills found in municipalities of Andalusia. First, we make a descriptive study of the variable. Second, we model this data according to the most suitable probabilistic model. Finally, several generalized linear regression models based on different geographic and socioeconomic variables are proposed and the best one (using the Akaike information criterion) is selected.

Keywords: Olive mills; Overdispersion; Count data; Regression models for count data; Goodness of fit.

JEL Classification: C13, R11

Received: November 19, 2021
Accepted: June 06, 2022

Sección Abierta

DOI: 10.25115/eea.v40i3.6927

Studies of Applied Economics / Estudios de Economía Aplicada

Volumen: 40-3, Septiembre 2022 // ISSN: 1133-3197

Una modelización del número de *almazaras* por municipio en Andalucía

VALENTINA CUEVA-LÓPEZ[1], JOSÉ RODRÍGUEZ-AVI[2], MARÍA JOSÉ OLMO-JIMÉNEZ[3], JULIA RODRÍGUEZ-REINOSO[4]

[1]Departamento de Estadística e I. O., UNIVERSITY OF JAÉN. SPAIN. E-mail: vcueva@ujaen.es
[2]Departamento de Estadística e I. O., UNIVERSITY OF JAÉN. SPAIN. E-mail: jravi@ujaen.es
[3]Departamento de Estadística e I. O., UNIVERSITY OF JAÉN. SPAIN. E-mail: mjolmo@ujaen.es
[4] E-mail: juliarodriguezreinoso@gmail.com

RESUMEN

Una *almazara* es una pieza esencial en la producción de aceite de oliva ya que es el lugar donde se muele la aceituna y se obtiene el aceite de oliva. Suelen estar vinculadas a cooperativas de productores. Son estructuras que requieren maquinaria especializada y que en múltiples ocasiones están infrautilizadas, dada la presencia de varias de ellas a muy corta distancia. Además, caracterizan a los municipios mayoritariamente olivareros y su estudio aporta una valiosa información de interés económico.

Desde el punto de vista estadístico, el "número de almazaras por municipio" es una variable de datos de conteo que presenta una sobredispersión. En este estudio, nos centramos en las almazaras existentes en los municipios de Andalucía. En primer lugar, realizamos un estudio descriptivo de la variable. En segundo lugar, modelizamos estos datos según el modelo probabilístico más adecuado. Finalmente, se proponen varios modelos de regresión lineal generalizada basados en diferentes variables geográficas y socioeconómicas y se selecciona el mejor (utilizando el criterio de información de Akaike).

Palabras clave: Almazaras; Sobredispersión; Datos de conteo; Modelos de regresión para datos de conteo; Bondad de ajuste.

Clasificación JEL: C13, R11

Recibido: 19 de Noviembre de 2021
Aceptado: 06 de Junio de 2022

1. Introduction

The cultivation of the olive tree and the production of olive oil and other related products (table olives, biomass, pharmaceutical products, etc.) is an increasingly important economic activity, since, although production is unevenly concentrated in 32 countries, the use of olive oil has spread almost all over the world. In particular, Spain produces around 50% of the olive oil produced in the world, and of that 50%, more than 60% is produced in Andalusia. Jaén, for its part, is the province in the world in which more oil is produced, with approximately 20% of world production. In addition, its production is a key element in the fixation of the population in the interior areas of many Spanish provinces.

Once the olive is harvested, it commences the production sector. There are four types of industries in this process. In the first place, the transformation of the olive into oil that is carried out in the *almazaras* (oil mills). The second step is packaging, which is carried out in the bottling machines, although there are oils with defects and they must be previously refined in the refineries. Finally, those that work with the olive pomace, a by-product of the passage through the oil mills and which are called extractors. Therefore, the first step in the production of the oil is carried out in the oil mills. According to AICA3 (2020), in Spain there are 1,826 active oil mills, distributed by 14 Autonomous Communities, and in greater numbers in Andalusia.

The oil mills are governed by business in two fundamental ways: cooperative (47.60%) and limited company (32.80%). According to Cooperativas Agroalimentarias de España (2017), of all olive oils produced in Spain, 70% are produced by cooperative oil mills. More recent data from AICA reduce this percentage to 66% (Parras-Mozas, 2021).

Consequently, the oil mill is a fundamental element of the producing areas. Four types of oil are extracted from them:

- *Extra virgin olive oil* (EVOO) which is characterized by having an acidity level of less than 0.8º. Its production process is more meticulous and its quality is much higher.
- *Virgin olive oil*, that has an acidity level of less than 1º and also contains highly valued monounsaturated fatty acids
- *Lampante or refined oil*, which is the one that has a lot of acidity and a very unpleasant taste and smell that prevents its consumption (in fact its name comes from its use as fuel in oil lamps).
- *Olive pomace oil,* a by-product obtained from the fat, residues of bones and skins of the olives that remain as remains of the oils.

The last two, especially lampante, are not directly consumable, but undergo suitable chemical processes, and are sold as refined olive oil (mostly lampante and with a mixture of between 10 and 20% EVOO) and olive pomace oil.

This classification (EVOO, virgin, refined and pomace) corresponds to the four commercial categories of olive oil recognized in the European Union legislation (EC Regulation 1019/2002). In consequence, knowledge and analysis of oil mills is an important aspect in the world of olive oil. In this work, we try to explain their number based on socioeconomic variables related to the municipalities in which they are located.

The rest of the paper is structured as follows: In Section 2 we include a descriptive summary of the variable and we model it using a univariate probabilistic model. In Section 3, we tackle the modelling of the variable using a generalized regression model, so we first describe the set of covariates used and finally we select the best regression model and the set of relevant covariates. In Section 4, we analyse the results provided by this regression model and, finally, in Section 5 we include a brief discussion about the whole work.

2. Study of the variable number of oil mills per municipality

In this work, a generalized regression model is considered in which the response or dependent variable Y is the number of oil mills per municipality in Andalusia in 2019. For this reason, first, we

include a descriptive analysis of this dataset and second, we model it using several count data distributions and we select the most appropriate.

2.1. Data description of the variable

The dependent variable in this study, number of oil mills per municipality in Andalusia in 2019, is a counting variable. The table of frequencies and the main statistics are shown in Table 1 and 2, respectively. It can be seen that the Aggregation Index $(AI = Var(Y) / \bar{Y})$ is $3.2368 > 1$, which indicates moderate overdispersion.

Table 1 Frequency table of the variable Y

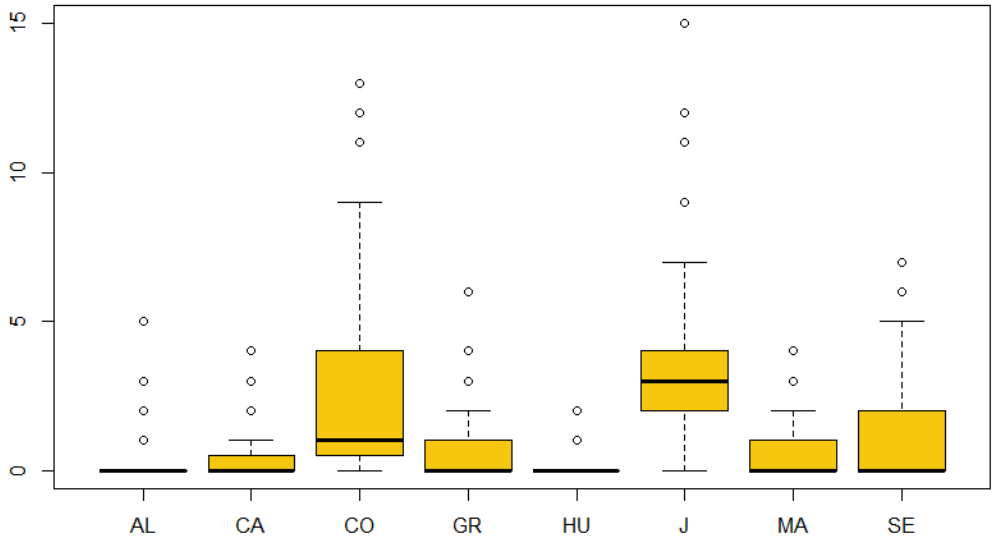
Y_i	0	1	2	3	4	5	6	7	8	9	11	12	13	15
n_i	412	159	93	48	21	14	10	6	3	5	3	2	1	1

Table 2 Descriptive summary of the variable Y

Mean	Variance	Q_1	Q_2	Q_3	Skewness	AI
1.1555	3.7403	0	0	2	2.9209	3.2368

Figure 1 shows the box-plot by provinces. We can observe a great variability between the provinces, highlighting Jaén and Córdoba. It should be pointed out the presence of relative extreme values within each province, even with atypical municipalities within them.

Figure 1 Box-plot of the variable Y grouped by provinces



2.2. Fit of a univariate probabilistic model

The variable Y is an overdispersed count-data variable, so we proposed the most common univariate probabilistic distribution to model this type of data. Specifically, we consider:

- The Complex Biparametric Pearson, CBP, distribution (Rodríguez-Avi et al, 2003 and Rodríguez-Avi & Olmo-Jiménez, 2017), with parameters $b > 0$ and $\gamma > 0$ and probability mass function (pmf) given by:

$$P(Y = y) = \frac{\Gamma(\gamma + bi)\Gamma(\gamma - bi)}{\Gamma(\gamma)^2} \frac{(bi)_y(-bi)_y}{(\gamma)_y} \frac{1}{y!}, \quad y = 0, 1, \ldots \tag{1}$$

  where $i = \sqrt{-1}$ is the imaginary unit and $(\alpha)_r = \Gamma(\alpha + r)/\Gamma(\alpha)$.

- The Complex Triparametric Pearson, CTP, distribution (Rodríguez-Avi et al, 2019), with parameters $\alpha \in \mathbb{R}$, $b, \gamma > 0$ and pmf given by:

$$P(Y=y)=\frac{\Gamma(\gamma-a+bi)\Gamma(\gamma-a-bi)}{\Gamma(\gamma)^2\Gamma(\gamma-2a)}\frac{(a+bi)_y(a-bi)_y}{(\gamma)_y}\frac{1}{y!},\quad y=0,1,\ldots \tag{2}$$

- The Negative Binomial distribution, NB, (Johnson et al, 2005), with parameters $\theta,\mu>0$ and pmf given by:

$$P(Y=y)=\frac{\Gamma(\theta+y)}{\Gamma(\theta)y!}\left(\frac{\theta}{\theta+\mu}\right)^{\vartheta}\left(\frac{\mu}{\theta+\mu}\right)^{y},\quad y=0,1,\ldots \tag{3}$$

- The Generalized Poisson distribution, GP, (Consul and Famoye, 1988) with parameters $\theta>0$ and λ, $\max\left(-1,-\frac{\theta}{m}\right)<\lambda<1$, and pmf given by:

$$P(Y=y)=\frac{\theta(\theta+\lambda y)^{y-1}}{y!}e^{-\theta-\lambda y},\quad y=0,1,\ldots \tag{4}$$

  when $\lambda>0$. If $\lambda<0$, $P(Y=y)=0$ for $y>m$, with $m\geq 4$ the largest positive integer for which $\theta+m\lambda>0$. This lower bound for λ is imposed to ensure at least 5 points in the sample space with positive probabilities. As a consequence of the definition, when $\lambda<0$ the distribution has a finite range (from 0 to m) and must be normalized, since the probabilities do not add up to 1. Trying to avoid this inconvenience, most packages in R (R Core Team, 2021) are restricted to the case $0<\lambda<1$.
- The Conway-Maxwell Poisson, CMP, distribution (Sellers et al, 2012), with $\lambda>0$, $\nu\geq 0$ and pmf given by:

$$P(Y=y)=\frac{\lambda^y}{(y!)^\nu}\frac{1}{Z(\lambda,\nu)},\quad y=0,1,\ldots \tag{5}$$

  with $Z(\lambda,\nu)=\sum_{j=0}^{\infty}\frac{\lambda^j}{(j!)^\nu}$.
- The Univariate Generalized Waring, UGW, distribution (Irwing, 1968; Xekalaki, 1983, Rodriguez-Avi et al, 2007), with parameters $a,k>0$ and $\rho>2$ and pmf given by:

$$P(Y=y)=\frac{\Gamma(a+\rho)\Gamma(k+\rho)}{\Gamma(a)\Gamma(k)\Gamma(\rho)}\frac{\Gamma(a+y)\Gamma(k+y)}{\Gamma(a+k+\rho+y)\Gamma(y+1)},\quad y=0,1,\ldots \tag{6}$$

- The Extended Bivariate Waring, EBW, distribution (Cueva-López et al, 2021) with parameters $\alpha\,\epsilon\,\mathbb{R},\gamma>\max(0,2\alpha)$ and pmf given by:

$$P(Y=y)=\frac{\Gamma(\gamma+\alpha)^2}{\Gamma(\alpha)^2\Gamma(\gamma-2\alpha)}\frac{\Gamma(\alpha+y)^2}{\Gamma(\gamma+y)\Gamma(y+1)},\quad y=0,1,\ldots \tag{7}$$

In order to select the best fit, a comparison is made according to two aspects. The χ^2 −goodness of fit test is used for checking which fitted model has the smallest distance with the observed data and the greater p-value. In addition, the Akaike Information Criterion, AIC, (Akaike, 1978) is calculated, which is defined as

$$AIC=-2\,ln\mathcal{L}(\boldsymbol{\theta})+2p \tag{8}$$

where $\mathcal{L}(\boldsymbol{\Theta})$ is the value of the likelihood function related to the fitted model and p the number of estimated parameters. According to this criterion, from Information theory, the best model is that with the lowest AIC. In all cases, the parameters are estimated by the maximum likelihood, ML, method. A summary of the estimates is shown in Table 3.

Table 3 χ^2 – goodness of fit test (statistic, degrees of freedom (df), p-value) and AIC for the fits of the variable Y

Distribution	AIC	Statistic	df	p-value
CTP	2292.822	4.952017	6	0.5499
CBP	2321.535	31.214837	7	<0.0001
GP	2287.665	3.547344	7	0.8302
EBW	2290.822	4.952047	7	0.6658
UGW	2292.822	4.852028	6	0.5499
CMP	2320.078	34.216267	5	<0.0001
NB	2289.244	4.944362	7	0.6668

As can be seen in Table 3, the best fit is that provided by the GP distribution, since it has the smallest distance (given by the χ^2 −test value), the greatest p −value of the test and the less AIC value. Other distributions for which the null hypothesis in the χ^2 −goodness of fit test is not rejected (also with p-values over 0.5) are the EBW, NB, CTP and UGW distributions.

Figure 2 shows the observed and expected values for each fit, which confirms the previous conclusion since the blue points (for the GP distribution) practically overlap the observed values.

Figure 2 Observed and expected frequencies of the different fits for the variable Y

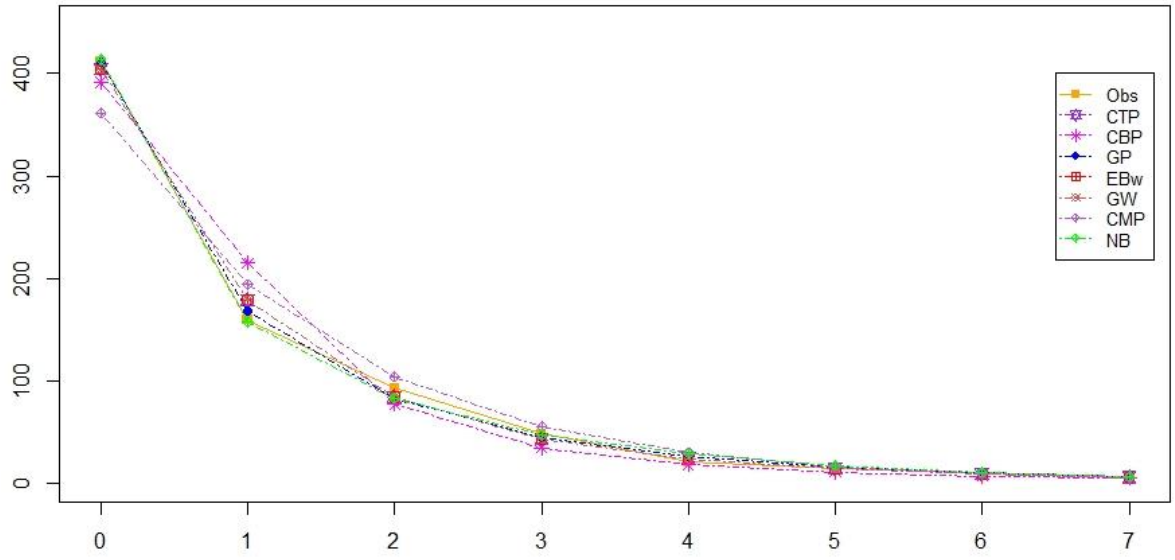


Table 4 includes the estimated parameters for the distributions that provide an appropriate fit according to the p-values in Table 3.

Table 4 ML parameter estimates and their corresponding standard errors (s.e.)

Model	Parameter	Estimate	s.e.
CTP	a	1.8134	0.1488
	b	0.0000	0.8987
	γ	7.4301	0.8244
GP	θ	0.6395	0.0329
	λ	0.4466	0.0264
EBW	λ	1.8135	0.1488
	ρ	3.8037	0.5382
UGW	a	1.8134	0.1517
	k	1.8194	0.1495
	ρ	3.8067	0.5382
NB	θ	0.5672	0.0552
	μ	1.1555	0.0672

3. Modelling through a regression model

3.1. Analysis of the selected covariates

Next, in order to find out what factors influence the number of oil mills per municipality, we model this variable using regression models based on Poisson, GP, NB and UGW distributions (Cameron-Trivedi, 2013; Hilbe, 2011; Famoye et al., 2004; Rodríguez-Avi et al., 2009).

The following variables have been considered as independent, explanatory or covariates, all of them measured in 2019 for each Andalusian municipality and obtained from the Multiterritorial Information System of Andalusia (SIMA) database (SIMA, 2021) of the Institute of Statistics and Cartography of Andalusia:

- Province to which each municipality belongs, with 8 categories (Almería, Cádiz, Córdoba, Granada, Huelva, Jaén, Málaga and Seville). It is coded by using 7 dummy variables with Jaén as base or reference category.
- Irrigated olive grove: Dummy variable that takes the value 1, if the main irrigated crop of the municipality is the olive grove and 0, otherwise.
- Dry olive grove: Dummy variable that takes the value 1, if the main dry land crop in the municipality is olive grove and 0, otherwise.
- % Under 20: Percentage of population under 20 years-old.
- % Over 65: Percentage of the population over 65 years-old.
- Mean age: Mean age in the municipality.
- Immigration × 1000 inhabitants: Internal, external and foreign immigrants per 1000 inhabitants.
- % Dry olive grove area: Percentage of dry olive grove cultivation area.
- % Irrigated olive grove: Percentage of irrigated olive grove cultivation area.
- Altitude: Altitude above sea level (in meters).
- % Men: Percentage of the male population.
- CSE: Number of Compulsory Secondary Education centers.
- Real estate × 1000 inhab.: Number of second-hand real estate transactions per 1000 inhabitants.

- Without employees × 1000 inhab.: Number of establishments without employees per 1000 inhabitants.
- % Men subsidized: Percentage of subsidized agricultural workers of the male sex.
- % Women subsidized: Percentage of subsidized female agricultural workers.
- Rustic IBI × 1000 inhab.: Number of rustic IBI per 1000 inhabitants.
- Unemployment rate: unemployment rate
- % Male unemployment: Percentage of registered unemployment in men.

A statistic summary of the quantitative covariates may be seen in Table 5.

Table 5 Statistic summary of quantitative covariates

Variable	Min	Max	Q_1	Me	Q_3	$\bar{x}$	s^2
% Under 20	1.70	30.00	14.50	18.20	21.00	17.68	21.71
% Over 65	6.60	46.60	16.80	20.60	25.57	21.47	41.82
Mean age	35.40	61.90	41.8	44.50	47.70	44.94	18.55
Immigration × 1000 hab.	4.31	172.73	26.42	38.31	57.20	44.89	653.75
% Dry olive grove area	0	91.11	0.81	4.64	19.14	15.58	430.48
% Irrigated olive grove	0	86.76	0.06	1.09	5.01	5.53	132.08
Altitude	2	1543	206	529.5	741.2	505.7	110785.2
% Men	46.27	60.71	49.66	50.38	51.50	50.79	3.26
CSE	0	48	0	1	1	1.437	10.52
Real state ×1000 inhab.	0	71.68	5.11	7.39	10.40	9.20	56.59
Without employees ×1000 inhab.	4.19	79.99	27.31	33.21	39.41	33.55	100.71
% Men subsidized	0	4.05	0.10	0.50	1.12	0.73	0.60
% Women subsidized	0	7.21	0.30	1.35	2.96	1.84	3.05
Rustic IBI × 1000 inhab.	0.13	4179.31	196.68	485.66	994.16	705.95	469891.3
Unemployment rate	6.90	38.71	17.83	20.68	23.78	20.95	24.36
% Male unemployment	1.24	15.33	5.07	6.47	8.02	6.63	4.73

3.2. Obtaining the regression model

These regression models have been estimated using R. Specifically, the `glm` function of the *stats* package has been used for the Poisson regression model, the `vglm` function of the *VGAM* package for the regression model based on the GP distribution, the `glm.nb` function from the *MASS* package for the NB regression model and the `gw` function from the *GWRM* package for the UGW regression model.

Table 6 contains the AIC values for each fitted regression model. The UGW-based regression model has been omitted because some convergence errors occur. The regression model that provides the best fit is the one based on the GP distribution.

Table 6 AIC values for the fitted regression models

Regression model	AIC
Poisson	1839.344
GP	1811.119
NB	1819.201

For the GP regression model, Table 7 includes the ML estimates of the regression coefficients, their standard errors and the p-values corresponding to the individual significance tests. Let us remember that in the GP regression model the parameter $\lambda \in (0,1)$ is fixed and the mean depends on the covariates through the expression (Consul and Famoye, 1992):

$$\mu_x = e^{\boldsymbol{x'\beta}} > 0 \tag{9}$$

where $x' = (1 \quad x_1 \quad \cdots \quad x_k)$ is the vector of covariates (including the constant) and $\boldsymbol{\beta} = (\beta_0 \quad \beta_1 \quad \cdots \quad \beta_k)'$ is the vector with the regression coefficients. Taking into account the expression of the distribution mean,

$$\mu = \frac{\lambda}{1-\theta}, \tag{10}$$

the parameter θ also depends finally on the covariates, $\theta_{\boldsymbol{x}}$. To guarantee that $\lambda \in (0,1)$ in the estimation process, a *logit* transformation is employed, so that:

$$\lambda_0 = \text{logit}(\lambda) = \ln\left(\frac{\lambda}{1-\lambda}\right) \Leftrightarrow \lambda = \frac{e^{\lambda_0}}{1+e^{\lambda_0}} \tag{11}$$

Table 7 ML regression coefficient estimates, standard error (s.e.) and p-values corresponding to the individual significance tests. GP regression model. * Indicates a statistically significant coefficient at 5%.

Parameters	Estimate	s.e.	p-value
(Intercept)	-2.5761	2.3893	0.2810
Province (Almería)	-0.8531*	0.2780	0.0021
Province (Cádiz)	-0.8388*	0.3047	0.0059
Province (Córdoba)	-0.0543	0.1333	0.6839
Province (Granada)	-0.7448*	0.1690	<0.0001
Province (Huelva)	-1.0097*	0.3007	0.0008
Province (Málaga)	-0.4186*	0.1774	0.0183
Province (Seville)	-0.5197*	0.1719	0.0025
Irrigated olive grove (Yes)	0.8758*	0.1610	<0.0001
Dry olive grove (Yes)	0.0506	0.1554	0.7445
% Under 20	0.0620*	0.0284	0.0290
% Over 65	0.0204	0.0212	0.3359
Immigration × 1000 hab.	-0.0099*	0.0034	0.0038
% Dry olive grove area	0.0144*	0.0020	<0.0001
% Irrigated olive grove	0.0120*	0.0028	<0.0001
Altitude	0.0004	0.0002	0.0701
% man population	-0.0343	0.0540	0.5245
CSE	0.0298*	0.0113	0.0084

Real state ×1000 inhab.	0.0082	0.0104	0.4260
Without employees ×1000 inhab.	0.0152*	0.0054	0.0046
% Men subsidized	0.0808	0.0973	0.4064
% Women subsidized	-0.0045	0.0390	0.9088
Rustic IBI × 1000 inhab.	-0.0389	0.0543	0.4740
Unemployment rate	-0.0002	0.0001	0.1038
% Male unemployment	0.0133	0.0220	0.5453
λ_0	-1.8317*	0.2385	<0.0001

Next, the explanatory variables of the model are selected applying stepwise selection procedures (forward, backward and combined) based on the AIC, so that a variable is included in the model if its presence increases it, while a variable is excluded from the model if its absence diminishes it. The model obtained in the backward stepwise regression differs slightly from those obtained in the forward and combined stepwise regression, presenting a lower AIC (1798.596 versus 1799.961). Then, we present the results of the former (Table 8). However, in both cases the AIC is reduced with respect to the full model. We observe that all the variables are significant at 10%, except for the dummy variable that represents the province of Córdoba, which indicates that there are no significant differences in the model for the number of oil mills per municipality between the provinces of Córdoba and Jaén.

Table 8 ML regression coefficient estimates, s.e. and p-values corresponding to the individual significance tests. GP stepwise regression model. * Indicates a statistically significant coefficient at 5%.

Parameters	**Estimate**	**s.e.**	**p-value**
(Intercept)	-1.7400*	0.4799	0.0003
Province (Almería)	-0.9347*	0.2548	0.0002
Province (Cádiz)	-0.9097*	0.2936	0.0019
Province (Córdoba)	-0.0555	0.1259	0.6594
Province (Granada)	-0.8112*	0.1511	<0.0001
Province (Huelva)	-1.0885*	0.2910	0.0002
Province (Málaga)	-0.4562*	0.1684	0.0068
Province (Sevilla)	-0.5647*	0.1650	0.0006
Altitude	0.0004	0.0002	0.0796
% under 20	0.0369*	0.0168	0.0279
Immigration ×1000 hab.	-0.0091*	0.0031	0.0032
CSE	0.0303*	0.0107	0.0045
% Dry olive grove area	0.0143*	0.0018	<0.0001
% Irrigated olive grove	0.0116*	0.0027	<0.0001
Irrigated olive grove (Yes)	0.8900*	0.1482	<0.0001
Without employees ×1000 inhab.	0.0154*	0.0051	0.0026
% men subsidized	0.1061	0.0601	0.0774
Rustic IBI × 1000 inhab.	-0.0002	0.0001	0.0757
λ_0	-1.8115*	0.2350	<0.0001

Based on the regression coefficient estimates of the dummy variables that represent the provinces, and their respective s.e., they can be grouped into three groups:

- First group: Almería, Cádiz and Huelva
- Second group: Granada, Malaga and Seville
- Third group: Córdoba and Jaén.

In this way, the regression model is fitted again, replacing the provinces by the aforementioned groups, considering the third (the provinces of Córdoba and Jaén) as the base or reference category. Thus, the results are shown in Table 9. This new fitted regression model has an AIC of 1793,913, lower than that of the fitted regression model with the disaggregated provinces. All the explanatory variables are relevant at 10%, except for *Altitude*, however, its exclusion from the model leads to a worse AIC.

Table 9 ML regression coefficient estimates, s.e. and p-values corresponding to the individual significance tests. GP stepwise regression model grouping the provinces. * Indicates a statistically significant coefficient at 5%.

Parameter	Estimate	s.e.	p-value
(Intercept)	-1.6686*	0.4477	0.0002
Group 1 (Almería+Cádiz+Huelva)	-0.9520*	0.1734	<0.0001
Group 2 (Granada+Málaga+Sevilla)	-0.6038*	0.1030	<0.0001
Altitude	0.0003	0.0002	0.1605
% under 20	0.0351*	0.0163	0.0317
Immigration ×1000 inhab.	-0.0097*	0.0028	0.0006
CSE	0.0307*	0.0106	0.0036
% Dry olive grove area	0.0148*	0.0017	<0.0001
% Irrigated olive grove	0.0122*	0.0025	<0.0001
Irrigated olive grove (Yes)	0.8482*	0.1425	<0.0001
Without employees ×1000 inhab.	0.0161*	0.0049	0.0010
% men subsidized	0.1025	0.0601	0.0880
Rustic IBI × 1000 inhab.	-0.0002	0.0001	0.0986
λ_0	-1.8021*	0.2334	<0.0001

4. Analysis of the results

We can point out the following interpretations from the fitted regression model selected:

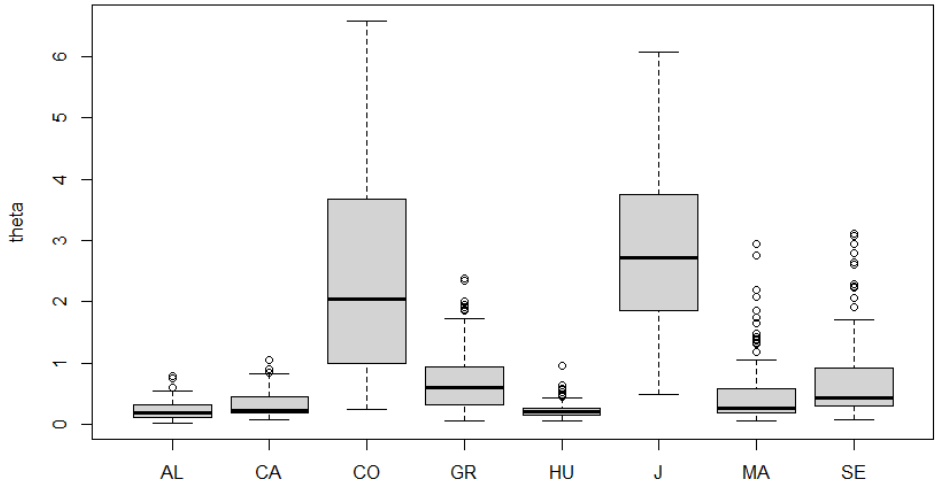
- The number of oil mills is a variable strongly related to the province. In this sense, the provinces of Jaén and Córdoba are those with the highest number, followed by the group of Granada, Málaga and Seville and lastly, the group composed by Almería, Jaén and Cádiz.
- The covariates that have a positive effect on the variable Y (they help to increase the average of Y) are the population under 20 years-old, the number of compulsory secondary education centers, the percentage of dry olive grove area, the percentage of irrigated olive grove, the presence of irrigated olive grove in the municipally, the number of establishments without salaried employees and the percentage of male subsidized workers.
- It should be considered that the fitted model provides an only estimate of the parameter λ, in this case 0.1408, and an estimate of θ for each combination of the covariates, (a vector of 778 values). A statistical summary of the ML estimates for θ is shown in Table 10.

Table 10 Statistical summary of θ −ML estimates

Mean	Variance	Min.	Q_1	Median	Mean	Q_3	Max.
0.9928	1.4701	0.0289	0.2155	0.4530	0.9928	1.3122	6.5714

The performance of θ is related to the values of the dependent variable Y. Figure 3 shows a box-plot of θ −ML estimates by provinces where a similar performance to that of Figure 1 can be observed. Furthermore, the Pearson correlation coefficient between the number of oil mills in the municipality and the estimate of the corresponding θ is 0.738.

Figure 3 Box-plot of $\boldsymbol{\theta}$ −ML estimates by provinces



To study the goodness of fit, a table of expected frequencies calculated from the fitted model can be obtained. To do this, the probability of the possible values of the variable is calculated for each residual distribution, $P[Y_i^* = k|GP(\hat{\theta}_i, \hat{\lambda})], k = 0, \ldots, 14$ and $P[Y_i^* \geq 15]$, with $i = 1, \ldots, 778$, where Y_i^* is the residual distribution provided by the fitted regression model for municipality i. Subsequently, the expected frequency provided by the model for each value from 0 to 15 is obtained by adding the probabilities in all municipalities. The results are shown in Table 11, where $\tilde{n}_i$ is the total expected frequency for $Y = i$.

Table 11 Observed and expected values of the variable Y

$\boldsymbol{Y}$	$\boldsymbol{n_i}$	$\widetilde{\boldsymbol{n}}_{\boldsymbol{i}}$
0	412	423.91
1	159	152.25
2	93	76.53
3	48	44.85
4	21	28.46
5	14	18.68
6	10	12.33
7	6	8.05
8	3	5.15
9	5	3.21
10	0	1.95
11	3	1.15
12	2	0.66
13	1	0.37
14	0	0.20
15	1	0.23

From Table 11, it is verified that

$$\frac{1}{778}\sum_{i=0}^{15} n_i Y_i \;=\; \frac{1}{778}\sum_{i=0}^{15} \tilde{n}_i Y_i \;= 1.5552 \tag{12}$$

which is the mean of the variable number of almazaras for municipality. Moreover, and given that R^2 is the proportion of variability of Y that is accounted for the covariates, a descriptive value of the goodness of fit is:

$$R^2 = \frac{\sum_{i=0}^{15} \tilde{n}_i (Y_i - 1.552)\text{^}2}{\sum_{i=0}^{15} n_i (Y_i - 1.552)\text{^}2} = \frac{2757.971}{2906.181} = 0.9490 \tag{13}$$

which is close to 1.

Additionally, this allows us to make interpretations for each municipality. Thus, from the estimates obtained in Table 9, and using equations (9) and (10), it is possible to know the residual distribution of each municipality based on the values of the covariates. In this way, the cumulative probability regarding the real number of oil mills can be calculated. As an example, Table 12 and 13 show these probabilities for the municipalities in the province of Jaén. Very high probabilities (e.g. over 90%) could indicate a saturation in that number in relation to what would be expected according to the characteristics of the municipality. In this situation are 13 of the 97 municipalities of the province (13.4%). The same, but in the opposite direction, could be said about the municipalities where this probability is very low (e.g. below 10%).

Table 12. Number of observed (O) and expected (E) oil mills and $\boldsymbol{P}\,(\boldsymbol{Y} \leq \boldsymbol{O})$ for the municipalities in the province of Jaén, based on the ML estimates of $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ provided by the fitted GP regression model in Table 9.

Municipality	$\boldsymbol{O}$	$\boldsymbol{E}$	$\boldsymbol{P}(\boldsymbol{Y} \leq \boldsymbol{O})$	Municipality	$\boldsymbol{O}$	$\boldsymbol{E}$	$\boldsymbol{P}(\boldsymbol{Y} \leq \boldsymbol{O})$
Albanchez de Mágina	3	1,490	0,910	Castillo de Locubín	3	2,518	0,742
Alcalá la Real	15	4,942	0,999	Cazalilla	1	2,690	0,298
Alcaudete	11	5,152	0,981	Cazorla	3	2,884	0,672
Aldeaquemada	1	0,779	0,810	Chiclana de Segura	2	0,977	0,900
Andújar	3	2,038	0,828	Chilluévar	1	5,830	0,036
Arjona	7	6,835	0,622	Escañuela	2	5,205	0,147
Arjonilla	3	6,024	0,191	Espeluy	1	1,962	0,457
Arquillos	3	2,678	0,712	Frailes	2	2,951	0,467
Arroyo del Ojanco	2	3,988	0,285	Fuensanta de Martos	5	5,095	0,604
Baeza	12	5,568	0,985	Fuerte del Rey	2	3,912	0,296
Bailén	5	5,278	0,576	Génave	2	2,339	0,601
Baños de la Encina	2	1,282	0,841	Guardia de Jaén, La	3	4,364	0,402
Beas de Segura	7	3,878	0,929	Guarromán	3	2,119	0,814
Bedmar y Garcíez	3	2,957	0,658	Higuera de Calatrava	4	5,605	0,378
Begíjar	2	4,184	0,258	Hinojares	0	0,572	0,612
Bélmez de la Moraleda	2	1,806	0,725	Hornos	1	1,234	0,665
Benatae	1	1,564	0,565	Huelma	5	3,172	0,869
Cabra del Santo Cristo	3	2,360	0,771	Huesa	1	2,229	0,392
Cambil	5	2,446	0,936	Ibros	3	3,640	0,528
Campillo de Arenas	2	2,324	0,604	Iruela, La	1	2,107	0,421
Canena	2	3,699	0,329	Iznatoraf	1	1,642	0,543
Carboneros	2	1,938	0,694	Jabalquinto	5	3,180	0,868
Cárcheles	2	3,222	0,414	Jaén	2	5,539	0,122
Carolina, La	2	2,154	0,644	Jamilena	2	4,482	0,220
Castellar	6	3,775	0,882	Jimena	2	3,801	0,313

Table 13. Table 12 (continuation)

Municipality	O	E	$P(Y \leq O)$	Municipality	O	E	$P(Y \leq O)$
Jódar	4	4,082	0,617	Santiago de Calatrava	2	3,636	0,340
Lahiguera	2	5,046	0,161	Santiago Pontones	0	1,365	0,310
Larva	1	2,757	0,286	Santisteban del Puerto	4	2,448	0,870
Linares	3	3,315	0,589	Santo Tomé	1	2,977	0,250
Lopera	2	4,962	0,169	Segura de la Sierra	3	1,907	0,849
Lupión	2	2,392	0,589	Siles	1	1,743	0,515
Mancha Real	9	5,531	0,916	Sorihuela del Guadalimar	2	1,849	0,715
Marmolejo	2	2,635	0,534	Torreblascopedro	3	3,393	0,574
Martos	9	6,134	0,876	Torredelcampo	6	6,537	0,534
Mengíbar	4	3,767	0,670	Torredonjimeno	9	7,073	0,796
Montizón	3	2,350	0,773	Torreperogil	2	5,110	0,156
Navas de San Juan	5	3,435	0,839	Torres	2	2,312	0,607
Noalejo	3	2,153	0,808	Torres de Albánchez	2	1,417	0,813
Orcera	2	1,755	0,737	Úbeda	11	5,167	0,981
Peal de Becerro	6	3,339	0,919	Valdepeñas de Jaén	3	2,880	0,673
Pegalajar	7	3,329	0,959	Vilches	4	1,844	0,935
Porcuna	6	6,670	0,516	Villacarrillo	5	4,128	0,747
Pozo Alcón	3	2,387	0,766	Villanueva de la Reina	6	1,801	0,989
Puente de Génave	2	4,646	0,201	Villanueva del Arzobispo	3	4,013	0,461
Puerta de Segura, La	4	3,054	0,785	Villardompardo	1	3,229	0,213
Quesada	3	2,227	0,795	Villares, Los	3	3,564	0,542
Rus	2	3,228	0,413	Villarrodrigo	2	1,054	0,886
Sabiote	3	4,937	0,317	Villatorres	5	4,941	0,627
Santa Elena	0	0,853	0,480				

5. Discussion

The study of oil mills within the economic sector has been the subject matter in various research works. In this work, another approach of the study of the number of oil mills is shown, emphasizing the different municipal variables that may influence whether or not there is an oil mill in the municipality. On the one hand, the procedure with which this approach has been carried out, involves proposing a univariate probabilistic model and then the relationship of this variable with a set of variables related to the municipality through a regression model for count data. In this way, a fitted model is obtained that allows the variables to be interpreted. Moreover, it provides a concrete and specific model based on the values that the explanatory variables present in each case. The number of explanatory variables, as well as their selection, may vary depending on the interest of the study and the availability of data, although the methodology developed here can be used in all cases.

The versatility of the procedure is noteworthy, since it can be applied to the independent variables considered appropriate by the researchers. The results also allow for a better understanding of the phenomenon studied, in this case the number of oil mills per municipality. It should be noticed the fact of being able to obtain probabilities for each municipality considering its own idiosyncrasy, obtaining the most appropriate distribution for the municipality according to the values of the significant covariates and determining the cumulative probability, which informs us about the level of scarcity or excess of that infrastructure. In particular, it can advise us on the convenience of expanding,

maintaining or reducing their number. This is a serious problem, since, in general, the oil mills are small and they are operational only during the period of olive harvesting and oil production, which must be done by allowing the shortest possible time between harvesting and grinding. An excess of them can increase production costs, due to the need to amortize the cost of the oil mills and it could be convenient, in certain cases, a concentration and reduction in their number. Of course, this decision must be adopted by other political and social reasons, but a data analysis as the one made in this work, can be helpful.

References

1. AICA3 (2020). Agencia de Información y Control Alimentarios. Información del sector. https://www.aica.gob.es/
2. Akaike, H. (1974), A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (6): 716-723.
3. Cameron A. C, Trivedi P. K. (2013). Regression analysis of count data, 2nd ed. Cambridge University Press, Cambridge
4. Consul, P. C.; Famoye, F. (1988): Maximum likelihood estimation for the generalized Poisson distribution when sample is larger than variance. Communications in Statistics: Theory and Methods, 17, 299-309.
5. Consul, P.C.; Famoye, F. (1992): Generalized Poisson Regression Model. Communications in Statistics: Theory and Methods, 21(1), 89-109.
6. Cueva-López, V.; Olmo-Jiménez, M.J.; Rodríguez-Avi, J. (2021). An over and underdispersed Biparametric extension of the Waring Distribution. Mathematics 9, 170.
7. Damas Rico, E. (1997). Análisis no paramétrico de la eficiencia relativa de las almazaras cooperativas en la provincia de Jaén durante el período 1975-1993. In Revista española de economía agraria (Issue 180, pp. 279–303). Ministerio de Agricultura, Pesca y Alimentación.
8. Famoye F, Wulu J. T., Singh K. P. (2004) On the generalized Poisson regression model with an application to accident data. J Sci 2:287–295
9. Hilbe J. M. (2011). Negative binomial regression, 2nd ed. Cambridge University Press, Cambridge
10. Irwing, J.O. (1968): The generalized Waring distribution applied to accident theory. Journal of the Royal Statistical Society. Series A, 131, 205-225.
11. Johnson, N. L.; Kemp, A. W.; Kotz, S. (2005): Univariate discrete distributions. Wiley, New York.
12. Ministerio de Agricultura, Pesca y Alimentación de España (2019): Encuesta sobre superficies y rendimientos de cultivo.
13. Olmo-Jiménez, M.J.; Rodríguez-Avi, J.; Cueva-López, V. (2019): A review of the CTP distribution: a comparison with other over- and underdispersed count data models. Journal of Statistical Computation and Simulation, 88(14), 2684-2706.
14. Parras Rosa, M and Mozas Moral, A. (2021). La Cadena de Valor de los Aceites de Oliva. In Informe anual de coyuntura del sector oleícola, pp 11-26. University of Jaén.
15. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
16. Rapoport, H.F. (2008): Botánica y morfología. En: Barranco, D., Fernández Escobar, R., Rallo, L., (Eds.). El cultivo del Olivo. 6ª Edición. Madrid. Junta de Andalucía y Ediciones Mundi-Prensa, 37-62.
17. Rodríguez-Avi, J.; Conde-Sánchez, A.; Sáez-Castillo, A. (2003): A new class of discrete distributions with complex parameters. Statistical Papers, 44, 67-88.
18. Rodríguez-Avi, J.; Conde-Sánchez, A.; Sáez-Castillo, A.; Olmo-Jiménez, M.J. (2007) A new generalization of the Waring distribution. Computational Statistics and Data Analysis, 51; 6138 – 6150
19. Rodríguez-Avi, J.; Conde-Sánchez, A.; Sáez-Castillo, A.; Olmo-Jiménez, M.J. Martínez-Rodríguez, A.M. (2009): A generalized waring regression model for count data. Comput Stat Data Anal, 53, 3717–3725.

20. Rodríguez-Avi, J.; Olmo-Jiménez, M.J. (2017): A regression model for overdispersed data without too many zeros. Statistical Papers, 58, 749-773.
21. Ruiz, C. (2006): Disfunciones en el gobierno de las sociedades cooperativas agrarias: el caso de las almazaras cooperativas. GEZKI, 2, 73-103.
22. Ruiz Jiménez, C.; García Martí, E.; Hernández Ortiz, M.J. (2013): Cómo responden a la crisis económica actual las Sociedades cooperativas agrarias. El caso de las Almazaras cooperativas andaluzas. REVESCO. Revista de Estudios Cooperativos, 113, 120-149.
23. Sellers, K.F.; Borle, S.; Shmueli, G. (2012). The COM-Poisson model for count data: a survey of methods and applications. Appl Stoch Models Bus Ind., 28, 104–116.
24. SIMA (2021). Sistema de Información Multiterritorial de Andalucía. IECA. Dirección web: https://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima/index2.htm
25. Xelakaki, E. (1983): The univariate generalized Waring distribution in relation to accident theory: proneness, spells or contagion? Biometrics, 39, 887-895.