PRONÓSTICO DE OCURRENCIA DE HELADAS MEDIANTE MODELOS DE CONTEOS INFLADOS POR CEROS

FROST INCIDENCE FORECAST USING ZERO INFLATED COUNT MODELS

C. Nelson Ariza Morales[1*], Paulino Pérez-Rodríguez[2]

Francisco Julián Ariza-Hernández[3]

[1]Departamento de Irrigación, Universidad Autónoma Chapingo, Carretera México-Texcoco km. 38.5, Chapingo, Estado de México, C. P. 56230, MÉXICO. Correo-e: nmariza15@gmail.com,

[2] Socio Economía Estadística e Informática, Colegio de Postgraduados, Texcoco, Estado de México, C.P. 56230, MÉXICO. Correo-e: perpdgo@colpos.mx,

[3]Facultad de Matemáticas, Universidad Autónoma de Guerrero, Chilpancingo, Guerrero, C.P. 39087, MÉXICO. Correo-e: arizahfj@uagro.mx.

(*Autor responsable).

ABSTRACT

In this paper we propose a statistical model to help us predict frost using climatological data such as temperature, relative humidity and precipitation. The problem will be addressed with the Quasi-Poisson, Hurdle and ZIP models. The models are fitted in the R software. The results of the study and the regression equation of the models are presented. Given the data, a comparison of the models is made using cross validation where we find that the ZIP model is the best model to predicts frost occurrence data.

Key words: Poisson, Hurdle, ZIP, R software, overdispersion.

RESUMEN

En este trabajo se propone un modelo estadístico que ayude a pronosticar la ocurrencia de heladas utilizando datos climatológicos como temperatura, humedad relativa y precipitación. El problema se abordará con los modelos de Quasi-Poisson, Hurdle y ZIP. El ajuste de los modelos se realiza utilizando el software R. Se presentan los resultados del estudio y la ecuación de regresión de los modelos. Dados los datos, se realiza una comparativa de los modelos usando la validación cruzada donde se encuentra que el modelo ZIP es el que mejor predice los datos de ocurrencia de heladas.

Palabras Clave: Poisson, Hurdle, ZIP, programa R, sobredispersión.

INTRODUCCIÓN

El clima está determinado por diversos factores, entre los cuales se encuentran la altitud, la latitud y la distribución existente de tierra y agua, solo por mencionar algunos. En el territorio nacional se identificaron 7 grandes tipos de climas, distribuidos a lo largo de todo el país. La gran diversidad de climas es factor importante para que haya diferentes fenómenos de origen meteorológico y que algunos de estos impactan a la población de manera directa o indirecta.

Es importante el estudio de la ocurrencia de heladas como fenómeno meteorologico debido al daño económico que puede ocasionar a nivel local y regional, por lo que el primer paso es obtener los datos de la temperatura mínima que definen la presencia o no de una helada y no es raro que al observar los datos el número de ceros en la variable respuesta sean los más representativos esto debido a que las heladas no ocurren con frecuencia en una zona pero al no prevenir una sola puede generar daños catastróficos, es por la razón anterior que para el estudio de este fenómeno es necesario recurrir a modelos estadísticos adecuados, es decir, a los modelos con exceso de ceros si es que se quiere pronosticar la ocurrencia de heladas. Actualmente se cuentan con diversos modelos de predicción de heladas con diferentes técnicas aplicadas al evento como: El caso de redes neuronales (Ovando, Bocco y Sayago 2003), método de mínimos cuadrados (Garcia 2018) y también con un modelo atmosférico (CONAGUA, Mapas de Pronóstico de Heladas Agrometeorológicas 2022). Aún con lo anterior, no se tienen reportes en la literatura del uso de modelos inflados por ceros para pronosticar la ocurrencia de heladas, y es otra razón importante que puede llevar a conclusiones importantes para esta aplicación.

En muchas aplicaciones de la ciencia se utilizan variables discretas o datos de conteos, debido a que es de interés conocer el número de veces que ocurre un evento en un periodo de tiempo determinado. Las variables discretas surgen en disciplinas tales como la demografía (Mamun 2014), la ingeniería (Xie, M. 2001), la biología (Thas y Rayner 2005), la agricultura (Ridout y Hinde 2001), entre otras. En cada una de las disciplinas anteriores y como muchas otras más están presentes los modelos inflados por ceros, por lo que cada vez es más imprescindible conocer y estudiar de manera objetiva este tipo de modelos.

Generalmente cuando se tiene que estudiar modelos para datos de conteos se suelen utilizar funciones como Poisson o Binomial Negativa. El modelo de regresión Poisson suele ser muy restrictivo para los datos de conteo debido a un fuerte supuesto de que la media y la varianza de los conteos son iguales y ocurre que el exceso de ceros en los conteos también conduce al problema de sobredispersión en el cual la varianza es mayor a la media, entonces Poisson subestima la probabilidad de que un conteo sea cero de lo que se observó en la muestra (Cameron y Trivedi 2005).

El interés de este estudio se centra en desarrollar y proponer un modelo probabilístico Poisson para la ocurrencia de heladas en la zona de Chapingo, Texcoco, Edo. de México. Los datos para el modelo propuesto estarán definidos por intervalos de una semana de forma de que los datos agroclimáticos obtenidos del Servicio Meteorológico Nacional y del portal Power (Prediction Of Worldwide Energy Resource, por sus siglas en inglés)[1] diseñado por National Aeronautics and Space Administration (NASA) estarán seccionados por 52 semanas del año. En este trabajo se utilizó el lenguaje de programación R (R Core Team 2022) y así como la librería de funciones pscl (Zeileis, Kleiber y Jackman 2008) (Political Science Computational Laboratory por sus siglas en inglés) la cual ofrece un método para el ajuste de modelos de regresión inflados con ceros para datos de conteos a utilizando el método de máxima verosimilitud.

METODOLOGÍA

Se obtuvo la temperatura mínima diaria de los años 1984 a 2009 de la estación Chapingo con clave 15170 y se construyó un registro de la ocurrencia o no de heladas, además en el mismo registro se añadieron los datos de temperatura mínima y máxima, humedad relativa y precipitación por ser variables que intervienen en la ocurrencia de heladas según la literatura. El registro generado fue con datos diarios por lo que se tomó la decisión para este estudio de trabajar con datos de ocurrencia de una semana, es decir cuántas ocurrencias de heladas teníamos en una

[1] (https://power.larc.nasa.gov/data-access-viewer/)

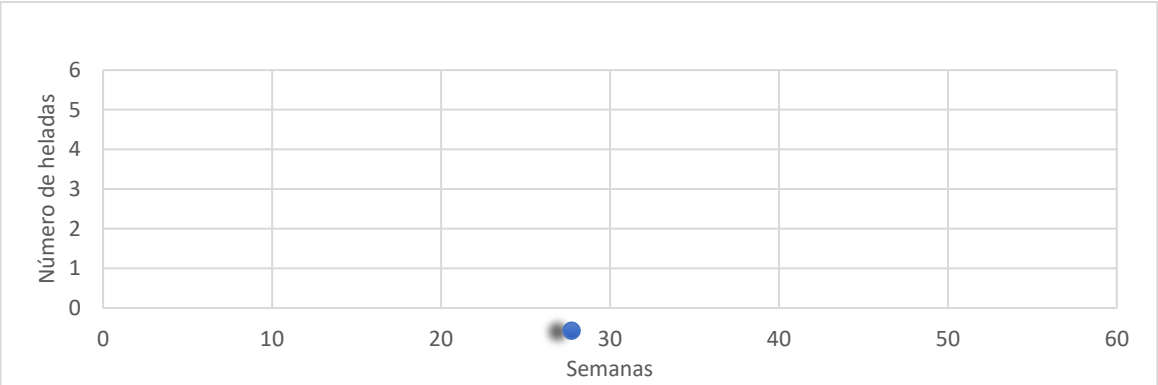


Figura 1. Eventos de heladas en la zona de Chapingo elaborados por los datos obtenidos del SMN de México para el año 2004. Fuente: Elaboración propia.

semana y así ocurrió para 52 semanas del año. Para el caso de las demás variables obtenidas se trabaja con rezagos, es decir, para fines de este trabajo 1 día antes de la ocurrencia de una helada, esto con el fin de pronosticar los eventos futuros.

En los datos históricos de ocurrencia de heladas se observó que los datos presentan un exceso de ceros, por lo que se tiene que recurrir a modelos inflados por ceros para trabajar con el pronóstico. En la Figura 1 se ilustra el problema que se discutió del exceso de ceros en los datos del estudio por lo que se pretende modelar esta ocurrencia de heladas con el modelo ZIP y el modelo Hurdle los cuales ayudan a abordar este problema de exceso de ceros, de igual forma se ajustó el modelo de Quasi-Poisson el cual según su literatura aborda la sobredispersión en los conteos de datos.

La ocurrencia de heladas se modela utilizando las covariables: Temperatura mínima y máxima, humedad relativa y la precipitación tomada previamente a que se presente el fenómeno de interés con motivos de prevenir el fenómeno.

Es primordial contar con todos los datos observados y tener fuentes confiables, ya que al momento de hacer el estudio son estos mismos datos los cuales ayudarán a crear el modelo y, además, a poder evaluar el mismo y tener la certeza de que el modelo funcionará con los datos históricos y a predecir datos futuros. En la Figura 1 se representa los datos observados de ocurrencia de heladas en Universidad Autónoma Chapingo, los cuales muestran gráficamente el exceso de ceros que se encontraron en el estudio. Es por lo anterior que para este estudio los datos provienen de dos sitios:

- Registros históricos del Sistema Meteorológico Nacional.
- Portal Power de la NASA.

Los datos provienen en esencia del SMN, pero para obtener datos faltantes y así mismo corroborar los datos históricos del SMN se utiliza el programa Power que ofrece la NASA.

Para este estudio se definirá la ocurrencia de heladas no solo por la temperatura mínima, sino que también se busca la relación que pueda tenerse con la temperatura máxima, humedad relativa y precipitación, es por lo que con la finalidad de tratar de entender de mejor manera el fenómeno de estudio y como está relacionado con las covariables por medio de las cuales se pretende predecir, es primordial hacer un estudio de covariables. Ahora, dado que las covariables en este estudio corresponden a variables medidas en escala fuerte se utiliza el coeficiente de correlación producto momento de Pearson.

Si la correlación es menor a cero, las variables se relacionan de forma inversa. Cuando el valor de alguna variable es alto, el valor de la otra variable es bajo. Mientras más próximo se encuentre a -1, más clara será la covariación extrema. Si el coeficiente es igual a -1, entonces se tiene una correlación negativa perfecta.

Cuando la correlación es igual a cero significa que no es posible determinar algún sentido de covariación. Sin embargo, no significa que no exista una relación no lineal entre las variables.

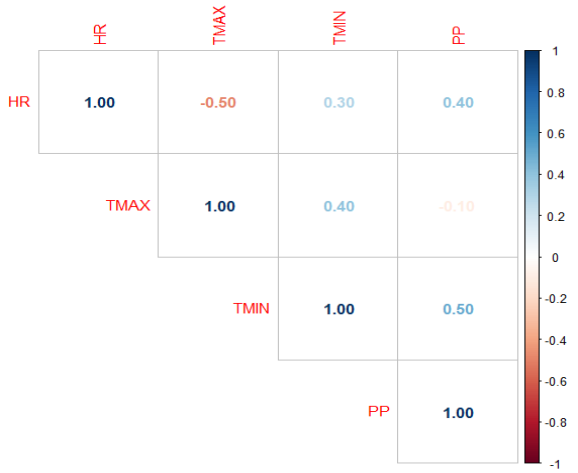


Figura 2. Coeficiente de Pearson para las covariables de HR, TMAX, TMIN y PP en donde se ven las relaciones inversas o directas

Ajuste de modelos de conteos inflados por ceros

Se debe usar modelos inflados por ceros debido a que el modelo Poisson no es recomendable por el exceso de ceros que tenemos en la muestra, razón por la cual se ajustaron los modelos Quasi-Poisson, Hurdle y Zero-Inflated Poisson con el fin de trabajar con la sobredispersión y con el exceso de ceros en nuestros datos.

Se analizaron los datos de 1352 semanas comprendidas de los años 1984-2009 de la estación Chapingo, Estado de México, México. Los datos se encuentran disponibles en la página del SMN en los archivos de Información Estadística Climatológica. El objetivo es modelar la ocurrencia de heladas según datos históricos de heladas durante las 1352 semanas.

Para el análisis de los datos se usó el programa R (R Core Team 2022), así como la biblioteca de funciones pscl (Zeileis, Kleiber y Jackman 2008).

RESULTADOS

Modelo Quasi-Poisson

Un modelo de regresión que puede manejar el problema de sobredispersión es la regresión de Quasi-Poisson (Hoef 2007) donde este modelo presta atención a la dispersión que provoca que la varianza de datos sea desigual a la media. Este modelo comienza en estimación de parámetros llamada cuasi-verosimilitud.

La función masa de probabilidades de una variable aleatoria Poisson con parámetro μ denotado por $Y \sim Poisson(\mu)$, está dado por:

$$f(y;\mu) = \frac{\exp(-\mu) \times \mu^{y}}{y!}, y = 0,1, \ldots$$

En el modelo Poisson $E(Y) = \mu$, mientras que en el modelo quasi Poisson, $E(Y) = \theta\mu$, donde θ es un parámetro de escala que ayuda a modelar la sobre-dispersión.

Utilizando la liga natural log, se tiene:

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots = \boldsymbol{x'\beta}$$

Donde x_1, x_2, x_3, x_4 corresponde a las covariables (HR, Tmax, Tmin, PP) y $\beta_0, \ldots, \beta_4$ corresponde a coeficientes de regresión a estimar.

Figura 3. Modelo Quasi-Poisson en comparación con los datos observados que, para fines de ilustrativos, la media del número de heladas predichas con este modelo se redondeó. Fuente: Elaboración propia.

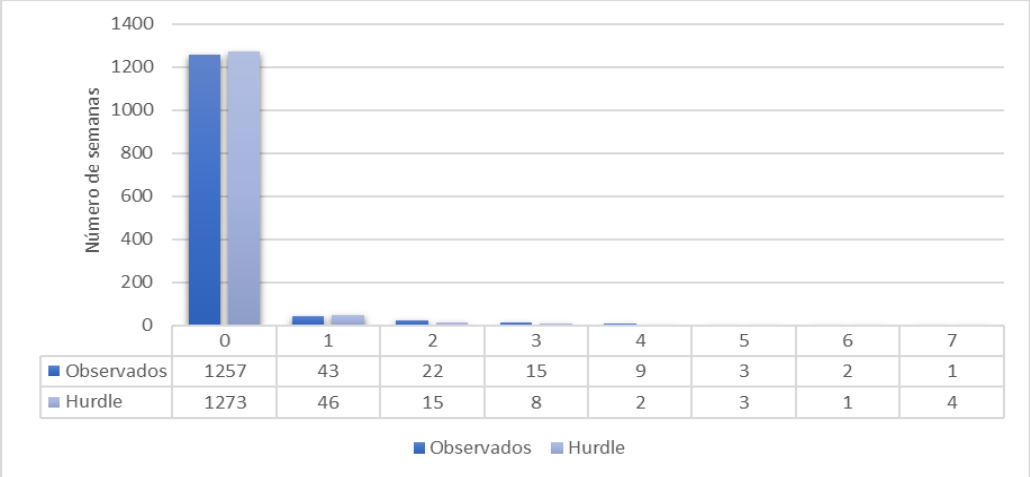


Figura 4. Resultados del modelo Hurdle junto con los datos observados para fines de ilustración los valores de la media se redondearon. Elaboración propia.

Modelo de Hurdle

El modelo de Hurdle combina un modelo de datos de recuento (que se trunca a la izquierda en y = 1) y un modelo de Hurdle cero (censurado a la derecha en y = 1): $f_{hurdle}(y; \mathbf{x}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}) =$

$$\begin{cases} f_{cero}(0, \boldsymbol{z}, \boldsymbol{\gamma}), & si\ y = 0 \\ \left(1 - f_{zero}(0, \boldsymbol{z}, \boldsymbol{\gamma})\right) \times \frac{f_{conteo}(y, \boldsymbol{x}, \boldsymbol{\beta})}{(1 - f_{conteo}(0, \boldsymbol{x}, \boldsymbol{\beta}))}, & si\ y > 0 \end{cases}$$

Utilizando la liga log se tiene:

$$\log(\mu) = \boldsymbol{x}'\boldsymbol{\beta} + \log\left(1 - f_{cero}(0, \boldsymbol{z}, \boldsymbol{\gamma})\right) - \log\left(1 - f_{conteo}(0, \boldsymbol{x}, \boldsymbol{\beta})\right)$$

Los parámetros de modelo $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ se estiman por máxima verosimilitud, donde la especificación de la probabilidad tiene la ventaja de que el recuento y el componente hurdle se pueden maximizar por separado. Así mismo el modelo anterior tiene dos regresores los cuales son los del modelo de conteo por ejemplo $y \sim x_1 + x_2$ en la notación usual de fórmulas en R y el modelo Hurdle por ceros ejemplo $y \sim z_1 + z_2$.

Modelos inflados por ceros (ZIP, Zero-Inflated Poisson)

La probabilidad de observar un recuento cero está en relación con la probabilidad:

$$\pi = f_{zero}(0; \boldsymbol{z}, \boldsymbol{\gamma})$$

$$f_{zeroinfl}(y; \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = f_{cero}(0; \boldsymbol{z}, \boldsymbol{\gamma}) \times I_{\{0\}}(y) + \left(1 - f_{cero}(0; \boldsymbol{z}, \boldsymbol{\gamma})\right) \times f_{conteo}(y; \boldsymbol{x}, \boldsymbol{\beta})$$

En donde $I_{\{0\}}$ es la función indicadora y la probabilidad no observada de pertenecer al componente de masa puntual se modela mediante un modelo lineal generalizado $\pi =$

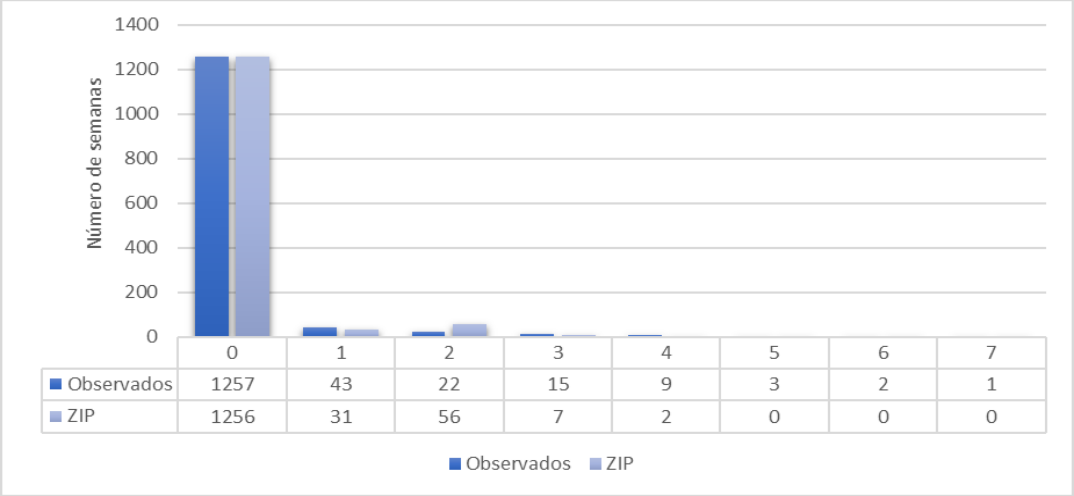


Figura 5. Modelo Zero- Inflated Poisson en comparación con los datos observados que para fines de ilustración la media del modelo ZIP se redondeó. Elaboración propia.

$g^{-1}(\boldsymbol{z}^{t}\boldsymbol{\gamma})$, donde g() corresponde a una función liga.

Y donde la ecuación de regresión correspondiente para la media es:

$$\mu = (1 - \pi) \times \exp(\boldsymbol{x}'\boldsymbol{\beta})$$

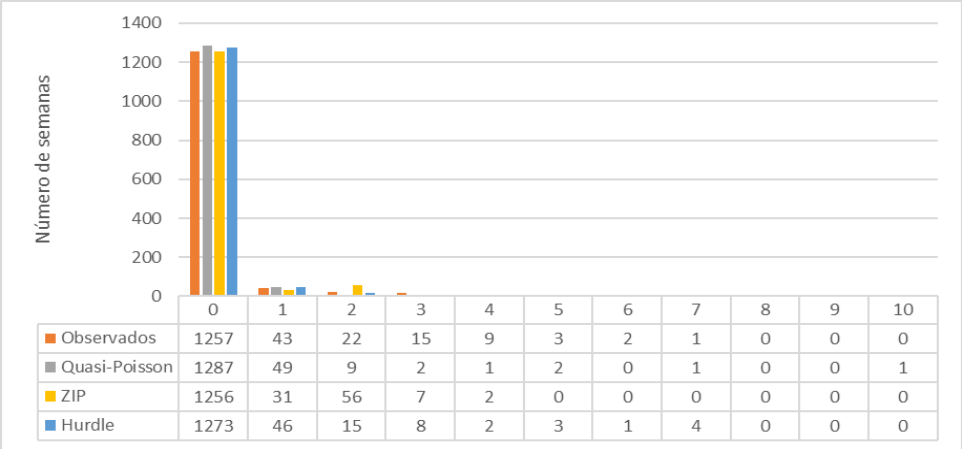
Comparación

En esta parte esencial se abordan las comparaciones entre los modelos ajustados. Uno de los primeros acercamientos a los resultados que daremos es la comparación entre las medias obtenidas de cada modelo, las cuales se representan en la Figura 6 de donde se observa que el que mejor describe los datos ZIP , ya que al comparar los valores observados vs predichos son bastante similares.

Existen otras maneras las cuales nos ayudan a comprobar el poder predictivo de un modelo. La validación es una fase importante del proceso de simulación que permite evaluar la calidad de un modelo. Específicamente en el caso de modelos de simulación continua se comparan datos y observaciones del sistema real con las predicciones generadas por el modelo.

Validación cruzada

Un problema recurrente en el ajuste de modelos es el sobreajuste, que implica que el modelo se ajusta muy bien a un conjunto de datos, pero puede no ser útil en el ajuste de otros datos, es decir, es menos generalizable. Una forma de validar el modelo es comprobar si el modelo predice correctamente un nuevo conjunto de datos.



	0	1	2	3	4	5	6	7	8	9	10
Observados	1257	43	22	15	9	3	2	1	0	0	0
Quasi-Poisson	1287	49	9	2	1	2	0	1	0	0	1
ZIP	1256	31	56	7	2	0	0	0	0	0	0
Hurdle	1273	46	15	8	2	3	1	4	0	0	0

Figura 6. Resultados de las medias que fueron redondeadas para fines prácticos de ilustración, comparando los tres modelos propuestos para conteos inflados por ceros.

El propósito principal de la validación cruzada es ver el comportamiento de un modelo con 80% de los datos trabajando con predicciones para el 20% de datos restante, esto para saber que tan bien predice con observaciones futuras de los datos restantes por lo que se realiza nuevamente el ajuste de todos los modelos, pero ahora con el 80% de los datos mencionados con la misma metodología presentada en este artículo.

Después de construir los modelos se tiene interés en determinar el poder predictivo de este para predecir el resultado de nuevas observaciones que no se usaron para construir el modelo. Es decir, queremos estimar el error de predicción.

Una estadística que ayuda a determinar la bondad de la predicción es la raíz cuadrada del cuadrado medio del error (RMSE por sus siglas en inglés), el cual mide el error de predicción promedio realizado por el modelo al predecir el resultado de una observación. Es decir, la diferencia promedio entre los valores de resultado conocidos observados y los valores predichos por el modelo. Cuanto menor sea el RMSE, mejor será el modelo.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

Modelo	RMSE
Quasi-Poisson	0.829
Hurdle	0.445
ZIP	0.300

Tabla 1. Resultados de RMSE para los datos de la validación cruzada con los datos de predicción.

Como se planteó anteriormente, el RMSE resulta que entre más bajo sea su valor el

modelo predice mejor los datos y en este caso corresponde al modelo ZIP. Con lo anterior se puede observar que la parte de los ceros estructurales no es significativa en los modelos, pero aun así al momento de predecir el evento tienen una muy buena cercanía mejor que el modelo Quasi-Poisson que trata la sobredispersión.

Es importante resaltar que cuando se hace el ajuste en los modelos de ZIP y Hurdle estos y se realiza la prueba de significancia de los coeficientes de regresión resultó que algunos coeficientes no son significativos, pero al momento de realizar la validación cruzada y comparar los resultados de los modelos se observa que el exceso de ceros es predicho de forma correcta tal como se muestra en las gráficas de valores observados vs predichos y en la validación cruzada con el RMSE.

CONCLUSIONES Y RECOMENDACIONES

El modelo Poisson, como fue mostrado en este trabajo, no resulta tan útil comparado con otros modelos para los conteos inflados por cero, pero la información que brinda puede servir de apoyo para un análisis previo y buscar el modelo inflado por ceros que más se ajuste.

En la literatura han sido utilizados diferentes modelos para el análisis de conteos inflados por ceros como la demografía (Mamun 2014), la ingeniería (Xie, M. 2001), la biología (Thas y Rayner 2005), la agricultura (Ridout y Hinde 2001), entre otros; sin embargo, no se tiene conocimiento del uso de modelos de regresión inflados por ceros en el estudio de la ocurrencia de heladas, por lo que resulta la metodología aplicada un buen acercamiento al pronóstico de heladas aplicable a diferentes zonas de estudios.

Dicho lo anterior y centrando la atención a los modelos ajustados inflados por ceros, resulta que, según las comparaciones y una validación cruzada trabajada, el modelo ZIP es el que mejor se ajusta a los datos, puesto que esta detectando toda esta ocurrencia de ceros en los datos de conteo y en las predicciones futuras.

Para la continuación de este estudio se recomienda hacer pruebas del trabajo con variables rezagadas a diversos grados, es decir, 2 días, 3 días, etc. Con el fin de comparar hasta qué punto o con que rezago podemos predecir el fenómeno de una helada ahora con el método de Poisson Inflado por ceros el cual resulto el mejor para este caso.

BIBLIOGRAFÍA

Cameron y Trivedi, 2005. *Microeconometrics: Methods an Applications.*. Cambrige University Press

CONAGUA, 2022. *Servicio Meteorológico Nacional.* https://smn.conagua.gob.mx/es/

Garcia, J., 2018. *Implementación de un modelo de pronóstico para el Valle del Mantaro.* Centro Internacional para la Investigación del Fenómeno de el Niño CIIFEN.

Hoef, J. a. B., 2007. *Quasi-Poisson Vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data?.* s.l.:Ecology.

Mamun, A., 2014. *Zero-Inflated regresion models for count data: an application to under- 5 deaths.* Ball State University

Ovando, G., Bocco, M. & Sayago, S., 2003. *Redes Neuronales para modelar la predicción de heladas.* Universidad Nacional de Córdoba, Facultad de Ciencias Agropecuarías, CC 509, 5000 Córdoba, Argentina

R Core Team, 2022. *R: A language and environment for statistical computing.* https://www.R-project.org/.

Ridout , M. y Hinde, J., 2001. *A score Test for testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives Biometrics.* University of Kent At Canterbury

Thas, O. y Rayner, J., 2005. *Smooth tests for the Zero-Inflated Poisson Distribution Biometrics.* Belgium: International

Xie, M., H., 2001. *Zero-Inflated Poisson model in statistical process control.* Singapore

Zeileis, A., Kleiber, C. y Jackman, S., 2008. *pscl.*https://www.jstatsoft.org/article/view/v027i08