

Multilabel prediction of cucurbit viruses from climatic variables using binary PLS and neural networks

Suarez Franco^{1,3} , Brugo Carivali Maria Florencia² , Perotto Maria Cecilia² 
and Bruno Cecilia^{1,3} 

¹ Grupo vinculado Estadística y Biometría a la Unidad de Fitopatología y Modelización Agrícola (UFyMA) – CONICET – INTA – Cordoba – Argentina.

² Unidad de Fitopatología y Modelización Agrícola (UFyMA) – CONICET – INTA – Cordoba – Argentina

³ Estadística y Biometría – Facultad de Ciencias Agropecuarias – Universidad Nacional de Córdoba. Córdoba – Argentina.

suarezfranco@agro.unc.edu.ar

Abstract. The performance of multilabel prediction models was evaluated for the detection of four non-persistently aphid-transmitted viruses in cucurbitaceae: watermelon mosaic virus (WMV), zucchini yellow mosaic virus (ZYMV), papaya ringspot virus (PRSV), and cucumber mosaic virus (CMV), using climatic variables. Partial least squares binary regression (PLSRBin) and neural networks (NN) were compared under both balanced and unbalanced class schemes. Model performance was assessed on a test dataset using metrics such as balanced accuracy, sensitivity, specificity, and the area under the ROC curve (AUC). Overall, NN model showed better performance compared to PLSRBin, particularly for CMV and PRSV, where the highest AUC and balanced accuracy values were observed. For WMV, both approaches yielded similar results, while ZYMV proved to be the most difficult virus to predict, with moderate performance across all models. The effect of class balancing varies depending on the virus and the modeling approach. In some cases, unbalanced models favor specificity at the expense of sensitivity, indicating a bias toward the majority class. ROC curves allowed for a clear visualization of differences between models, showing a consistent advantage of NN in most scenarios. Overall, the results suggest that flexible models such as NN are better suited to capture the complexity of the data, although the choice of modeling approach and class balancing strategy should be guided by the specific objectives of the analysis.

Keywords: prediction model, artificial networks, plant disease.

Predicción multietiqueta de virus de cucurbitáceas a partir de variables climáticas mediante PLS binario y redes neuronales

Resumen. Se evaluó el desempeño de modelos multirrespuesta para la detección de la presencia/ausencia de cuatro virus transmitidos por áfidos de manera no persistente en cucurbitáceas: watermelon mosaic virus (WMV), zucchini yellow mosaic virus (ZYMV), papaya ringspot virus (PRSV) y cucumber mosaic virus (CMV) a partir de variables climáticas, comparando regresión por mínimos cuadrados parciales binaria (PLSRBin) y redes neuronales (NN), con y sin balanceo de clases. La evaluación se realizó sobre un conjunto de prueba utilizando métricas como: precisión balanceada, sensibilidad, especificidad y área bajo la curva ROC (AUC). En términos generales, los modelos de NN mostraron un mejor desempeño respecto a PLSRBin, particularmente para CMV y PRSV, donde se observaron los mayores valores de AUC y precisión balanceada. En el caso de WMV, ambos enfoques presentaron rendimientos similares, mientras que ZYMV resultó ser el virus más difícil de predecir, con valores de desempeño moderados en todos los modelos. El efecto del balanceo fue variable según el virus y el modelo. En algunos casos, el no balanceo favoreció la especificidad a costa de una menor sensibilidad, evidenciando un sesgo hacia la clase mayoritaria. Las curvas ROC permitieron visualizar claramente las diferencias entre modelos, mostrando una ventaja consistente de las NN en la mayoría de los escenarios. En conjunto, los resultados sugieren que modelos flexibles como las NN capturan mejor la complejidad de los datos, aunque la elección del enfoque y el tratamiento del desbalance deben considerarse en función del objetivo específico del análisis.

Palabras clave: modelos de predicción, red artificial, enfermedades vegetales.

1 Introducción

La predicción de ocurrencia de enfermedades en cultivos agrícolas constituye un desafío estadístico-metodológico debido a la naturaleza multicausal de los procesos epidemiológicos. Entre ellos, el desbalance de clase (presencia/ausencia), la co-presencia de enfermedades y la posible relación no lineal entre variables ambientales y ocurrencia de enfermedad. En cucurbitáceas, *watermelon mosaic virus* (WMV), *zucchini yellow mosaic virus* (ZYMV), *papaya ringspot virus* (PRSV) y *cucumber mosaic virus* (CMV) son los virus no persistentes, transmitidos por áfidos, de mayor incidencia y distribución en los principales lotes productivos de Argentina. Su detección temprana resulta clave para orientar decisiones de manejo fitosanitario, dado que la eficacia del manejo químico de los vectores es limitada frente a virus transmitidos de manera no persistente (Zaffaroni et al., 2021). La disponibilidad de productos de reanálisis climático junto con el desarrollo de métodos de aprendizaje automático ofrecen nuevas oportunidades para construir modelos predictivos (Paccioretti et al., 2023). El objetivo de este trabajo fue evaluar la capacidad predictiva de variables climáticas derivadas de ERA5-Land para anticipar la presencia/ausencia simultánea de WMV, ZYMV, PRSV y CMV en lotes productivos de cucurbitáceas de Argentina. Para ello, se compararon dos enfoques de modelado multietiqueta con diferente grado de flexibilidad estructural: una red neuronal artificial (NN), y un modelo de regresión por mínimos cuadrados parciales para respuestas binarias multivariadas (PLSRBin). Además, se evaluó el efecto del balanceo de clases sobre el desempeño predictivo de ambos enfoques.

2 Materiales y métodos

2.1 Datos epidemiológicos

Los datos de presencia/ausencia de virus fueron relevados de lotes productivos de cucurbitáceas entre 2010 y 2024 en 14 provincias Argentina (Noroeste, Noreste, Centro y Cuyo). Cada campaña, los lotes seleccionados fueron inspeccionados visualmente para identificar plantas con síntomas compatibles con infecciones virales. Se recolectaron muestras de tejido joven sintomático y la detección de WMV, ZYMV, PRSV y CMV se realizó mediante DAS ELISA. Se construyó una base de datos con 1.286 registros y una variable respuesta binaria para cada virus.

2.2 Datos climáticos

Se descargaron variables meteorológicas horarias considerando 90 días desde la fecha de siembra con el paquete *ecmwfr* de R (Hufkens et al., 2019) a partir del producto de reanálisis ERA5-Land Hourly Averaged (Muñoz-Sabater et al., 2021). Las variables descargadas incluyeron temperatura media, velocidad media del viento, temperatura media del punto de rocío, precipitación y evaporación acumulada. La humedad relativa

se estimó a partir de la temperatura del aire y la temperatura del punto de rocío (Lawrence, 2005) y la velocidad del viento a partir de sus componentes u y v propuesto en ERA5.

2.3 Preprocesamiento

Se trabajó con la presencia/ausencia de cada virus y se construyó un perfil multiviral de cada observación utilizado para particionar el conjunto de datos en entrenamiento (80%) y prueba (20%). Este procedimiento permitió preservar la distribución conjunta de las respuestas virales en ambos subconjuntos. Dado el desbalance entre clases de presencia/ausencia, se aplicó un procedimiento de sobremuestreo con reposición sobre el subconjunto de entrenamiento, balanceando secuencialmente cada virus desde el de mayor desbalance hasta igualar la frecuencia de ambas clases. El conjunto de prueba se mantuvo sin modificaciones para evitar sesgos en la evaluación del modelo.

2.4 Modelado predictivo

Se construyeron dos esquemas de entrenamiento para la predicción simultánea de la presencia/ausencia de WMV, ZYMV, PRSV y CMV: uno utilizando el conjunto original (sin balanceo) y otro el balanceado. En ambos casos, las variables de respuesta fueron modeladas de manera conjunta bajo un enfoque de clasificación multietiqueta. Se evaluaron dos algoritmos de modelado: **Red neuronal artificial (NN)**: Las variables predictoras fueron la matriz de entrada y las cuatro respuestas binarias virales la de salida. La arquitectura consistió en una red secuencial hacia adelante (feed-forward) con tres capas ocultas densamente conectadas de 32, 16 y 8 neuronas, respectivamente, todas con función de activación ReLU, seguidas de una capa de salida de cuatro neuronas con activación sigmoide, una por cada virus. El modelo se compiló utilizando el optimizador Adam (tasa de aprendizaje de 0.001) y función de pérdida binary cross-entropy. El entrenamiento se realizó durante 150 épocas como máximo, con tamaño de lote 32 y una partición interna para validación del 20%. Para reducir el riesgo de sobreajuste, se implementó early stopping. La red se implementó en R, paquete keras3 (Kalinowski et al., 2025). **Regresión por mínimos cuadrados parciales binaria (PLSRBin)**: La matriz de predictores constituyó el bloque explicativo y las cuatro respuestas virales el bloque respuesta. PLSRBin modela simultáneamente múltiples variables binarias correlacionadas mediante componentes latentes. El modelo se ajustó con siete componentes latentes y un parámetro de penalización de 0.1. Previo al ajuste, las variables predictoras fueron estandarizadas. Se obtuvieron probabilidades predichas para cada virus, transformadas a clases binarias con umbral de decisión de 0.5. La implementación se realizó con el paquete MultiBplotR (Vicente-Villardón et al., 2023).

2.5 Evaluación del desempeño predictivo

Para cada virus, en el conjunto de prueba, se estimó precisión, sensibilidad, especificidad, precisión balanceada y área bajo la curva ROC (AUC). Además, se construyeron

curvas ROC para cada virus con el fin de comparar la capacidad discriminatoria de los modelos y evaluar el efecto del balanceo de clases sobre el rendimiento predictivo.

3 Resultados

Las NN mostraron un desempeño consistentemente superior a los modelos PLSRBin. En términos generales, los mayores valores de precisión balanceada (PB) (Tabla 1) y AUC (Fig. 1) se obtuvieron para WMV y CMV, mientras que ZYMV y en menor medida PRSV, presentaron menor capacidad predictiva. El mejor desempeño global se observó para CMV con NN balanceada (PB=0.811; AUC=0.915), seguido por WMV con el mismo enfoque (PB=0.819; AUC=0.881). Para PRSV, la NN también superó a PLSRBin, alcanzando valores de AUC de 0.815 y 0.805 en las versiones balanceada y no balanceada, respectivamente. En cambio, ZYMV fue la respuesta más difícil de predecir, con los menores valores generales de precisión balanceada, particularmente bajo PLSRBin. El balanceo de clases tendió a mejorar el desempeño de las NN, especialmente en términos de precisión balanceada, con mejoras evidentes en CMV, ZYMV y WMV. Este efecto fue particularmente marcado en los virus con mayor desbalance original: CMV presentó la mayor proporción de ausencias en el conjunto sin balancear (89,0%), seguido por PRSV (79,7%), ZYMV (62,2%) y WMV (50,1%). Consistentemente, en CMV sin balanceo la NN alcanzó una especificidad de 0.919 pero una sensibilidad de solo 0.579, reflejando un sesgo hacia la clase mayoritaria que el balanceo corrigió (sensibilidad=0.895; especificidad=0.828). En conjunto, las curvas ROC confirmaron una mayor capacidad discriminatoria de las NN respecto a PLSRBin, sugiriendo que la relación entre las variables climáticas y la ocurrencia viral fue mejor capturada mediante enfoques no lineales.

Tabla 1. Precisión balanceada, sensibilidad, especificidad y Area bajo la curva (AUC) de los modelos evaluados para la detección de cuatro virus de cucurbitácea, con y sin balance.

Modelo	Balanceado	Virus	Sens.	Esp.	Prec. balanceada	AUC
Red neuronal artificial	Sí	WMV	0.775	0.874	0.819	0.868
		ZYMV	0.539	0.829	0.696	0.746
		PRSV	0.667	0.898	0.781	0.801
		CMV	0.895	0.828	0.811	0.940
	No	WMV	0.775	0.883	0.791	0.879
		ZYMV	0.645	0.775	0.628	0.766
		PRSV	0.615	0.922	0.774	0.809
		CMV	0.579	0.919	0.728	0.930
Regresión por Mínimos Cuadrados Parciales Binaria	Sí	WMV	0.765	0.748	0.756	0.788
		ZYMV	0.658	0.527	0.593	0.665
		PRSV	0.513	0.723	0.618	0.629
		CMV	0.789	0.688	0.739	0.798
	No	WMV	0.767	0.720	0.743	0.824
		ZYMV	0.623	0.710	0.667	0.709
		PRSV	0.378	0.818	0.598	0.700
		CMV	0.258	0.909	0.819	0.732

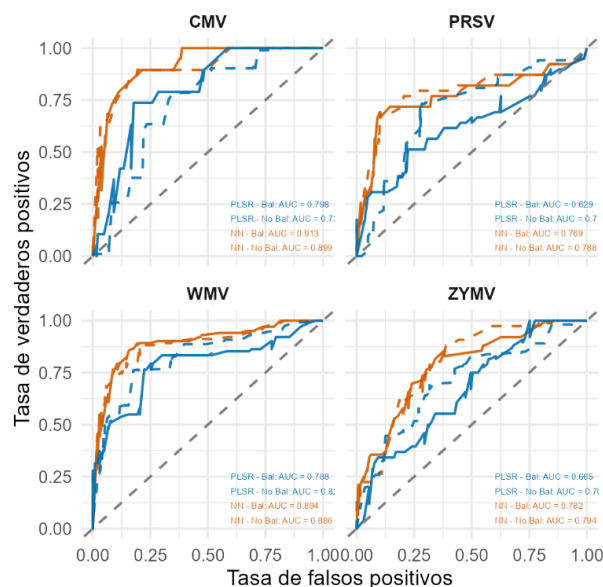


Fig. 1. Curvas ROC para cada virus (WMV, ZYMV, PRSV y CMV) bajo modelos de red neuronal artificial (NN, naranja) y Regresión por mínimos cuadrados parciales binaria (PLSR-Bin, azul) con balanceo (línea continua) y sin balanceo (línea de puntos) de clases.

References

- Hufkens, K., Stauffer, R., & Campitelli, E. (2019). *The ecwmfr package: an interface to ECMWF API endpoints*.
- Kalinowski, T., Allaire, J. J., & Chollet, F. (2025). *keras3: R Interface to "Keras."*
- Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J. N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9), 4349–4383.
- Paccioretti, P., Giannini-Kurina, F., Suarez, F., & Scavuzzo, M. (2023). Protocolo para automatizar la descarga de datos climáticos desde la nube y generar indicadores biometeorológicos para el monitoreo epidemiológico de cultivos. *AgriScientia*, 1(40 SE-Comunicaciones).
- Vicente-Villardón, J. L., Vicente-Gonzalez, L., & Frutos-Bernal, E. (2023). *MultiBiplotR: Multivariate Analysis Using Biplots in R*.
- Zaffaroni M, Rimbaud L, Mailleret L, Cuniffe NJ, Bevacqua D. Modelling Interference Between Vectors of Non-Persistent and Transmitted Plant Viruses to Identify Effective Control Strategies. *PLOS Computational Biology*. 2021; 17(12).