

Estratificación temporal de Aedes Aegypti basada en herramientas geoespaciales y aprendizaje automático

Juan M. Scavuzzo

Facultad de Matemática, Astronomía, Física y Computación
Universidad Nacional de Córdoba

Diciembre de 2018

Motivación de este trabajo: problemática epidemiológica

- Aedes aegypti es el principal vector de Dengue, Chikungunya, Zika y Fiebre Amarilla urbana

Motivación de este trabajo: problemática epidemiológica

- Aedes aegypti es el principal vector de Dengue, Chikungunya, Zika y Fiebre Amarilla urbana
- Datos de la Organización Mundial de la Salud:

Motivación de este trabajo: problemática epidemiológica

- Aedes aegypti es el principal vector de Dengue, Chikungunya, Zika y Fiebre Amarilla urbana
- Datos de la Organización Mundial de la Salud:
 - 80 millones de personas se infectan de Dengue anualmente

Motivación de este trabajo: problemática epidemiológica

- Aedes aegypti es el principal vector de Dengue, Chikungunya, Zika y Fiebre Amarilla urbana
- Datos de la Organización Mundial de la Salud:
 - 80 millones de personas se infectan de Dengue anualmente
 - 550 mil enfermos requieren hospitalización

Motivación de este trabajo: problemática epidemiológica

- Aedes aegypti es el principal vector de Dengue, Chikungunya, Zika y Fiebre Amarilla urbana
- Datos de la Organización Mundial de la Salud:
 - 80 millones de personas se infectan de Dengue anualmente
 - 550 mil enfermos requieren hospitalización
 - 20 mil personas mueren

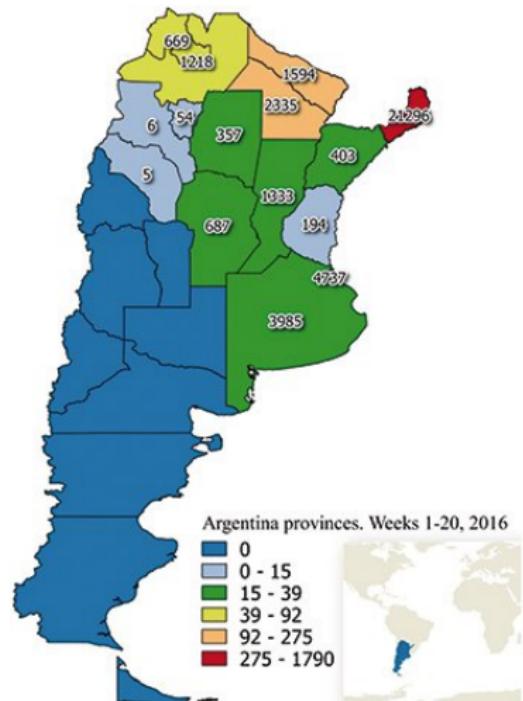
Motivación de este trabajo: problemática epidemiológica

- Aedes aegypti es el principal vector de Dengue, Chikungunya, Zika y Fiebre Amarilla urbana
- Datos de la Organización Mundial de la Salud:
 - 80 millones de personas se infectan de Dengue anualmente
 - 550 mil enfermos requieren hospitalización
 - 20 mil personas mueren
 - 2.500 millones de personas corren riesgo de contraer la enfermedad

Motivación de este trabajo: problemática epidemiológica

- Aedes aegypti es el principal vector de Dengue, Chikungunya, Zika y Fiebre Amarilla urbana
- Datos de la Organización Mundial de la Salud:
 - 80 millones de personas se infectan de Dengue anualmente
 - 550 mil enfermos requieren hospitalización
 - 20 mil personas mueren
 - 2.500 millones de personas corren riesgo de contraer la enfermedad
 - Más de 100 países con transmisión endémica

Motivación de este trabajo: problemática epidemiológica



Motivación de este trabajo: problemática epidemiológica

Características del Aedes aegypti

Motivación de este trabajo: problemática epidemiológica

Características del Aedes aegypti

- Gran capacidad adaptativa

Motivación de este trabajo: problemática epidemiológica

Características del Aedes aegypti

- Gran capacidad adaptativa
- Resistencia a insecticidas

Motivación de este trabajo: problemática epidemiológica

Características del Aedes aegypti

- Gran capacidad adaptativa
- Resistencia a insecticidas
- Resistencia de huevos a la desecación

Motivación de este trabajo: problemática epidemiológica

Características del Aedes aegypti

- Gran capacidad adaptativa
- Resistencia a insecticidas
- Resistencia de huevos a la desecación
- Presencia en el medio urbano

Motivación de este trabajo: problemática epidemiológica

Características del Aedes aegypti

- Gran capacidad adaptativa
- Resistencia a insecticidas
- Resistencia de huevos a la desecación
- Presencia en el medio urbano
- Preferencia de cría en contenedores artificiales

Motivación de este trabajo: sistemas de modelado actuales

Modelar utilizando información satelital

Motivación de este trabajo: sistemas de modelado actuales

Modelar utilizando información satelital

- Información ambiental con alcance regional

Motivación de este trabajo: sistemas de modelado actuales

Modelar utilizando información satelital

- Información ambiental con alcance regional
- Información espacio-temporal

Motivación de este trabajo: sistemas de modelado actuales

Modelar utilizando información satelital

- Información ambiental con alcance regional
- Información espacio-temporal
- Grandes avances en los últimos años

Motivación de este trabajo: sistemas de modelado actuales

Modelar utilizando información satelital

- Información ambiental con alcance regional
- Información espacio-temporal
- Grandes avances en los últimos años

Motivación de este trabajo: sistemas de modelado actuales

Modelar utilizando información satelital

- Información ambiental con alcance regional
- Información espacio-temporal
- Grandes avances en los últimos años

Pero actualmente se utilizan modelos lineales para relacionar las distintas variables!

Motivación de este trabajo

Algunas cuestiones a tener en cuenta

Motivación de este trabajo

Algunas cuestiones a tener en cuenta

- La prevención de las enfermedades en cuestión debe ser a través de control de vectores

Motivación de este trabajo

Algunas cuestiones a tener en cuenta

- La prevención de las enfermedades en cuestión debe ser a través de control de vectores
 - Modelos predictivos!

Motivación de este trabajo

Algunas cuestiones a tener en cuenta

- La prevención de las enfermedades en cuestión debe ser a través de control de vectores
 - Modelos predictivos!
- Será correcto asumir relaciones lineales entre las variables ambientales y la abundancia del vector?

Motivación de este trabajo

Algunas cuestiones a tener en cuenta

- La prevención de las enfermedades en cuestión debe ser a través de control de vectores
 - Modelos predictivos!
- Será correcto asumir relaciones lineales entre las variables ambientales y la abundancia del vector?
 - Modelos no-lineales con... aprendizaje automático!

Motivación de este trabajo

Algunas cuestiones a tener en cuenta

- La prevención de las enfermedades en cuestión debe ser a través de control de vectores
 - Modelos predictivos!
- Será correcto asumir relaciones lineales entre las variables ambientales y la abundancia del vector?
 - Modelos no-lineales con... aprendizaje automático!
- Si es un sistema regional, cómo extrapoló los modelos?

Motivación de este trabajo

Algunas cuestiones a tener en cuenta

- La prevención de las enfermedades en cuestión debe ser a través de control de vectores
 - Modelos predictivos!
- Será correcto asumir relaciones lineales entre las variables ambientales y la abundancia del vector?
 - Modelos no-lineales con... aprendizaje automático!
- Si es un sistema regional, cómo extrapoló los modelos?
 - A través de relaciones entre características ambientales!

Objetivos

Objetivos

- Implementar una herramienta, sencilla, para generar modelos predictivos

Objetivos

- Implementar una herramienta, sencilla, para generar modelos predictivos
- Validar la hipótesis de que "modelos no-lineales son mejores para predecir la oviposición que los lineales"

Objetivos

- Implementar una herramienta, sencilla, para generar modelos predictivos
- Validar la hipótesis de que "modelos no-lineales son mejores para predecir la oviposición que los lineales"
- Proponer una solución a la problemática de escases de datos que se evidencia al pensar en sistemas regionales de estimación de riesgo

Algunos Conceptos: *Epidemiología Panorámica*

Algunos conceptos importantes

Algunos Conceptos: *Epidemiología Panorámica*

Algunos Conceptos: *Epidemiología Panorámica*

- La teledetección y su capacidad de adquirir información

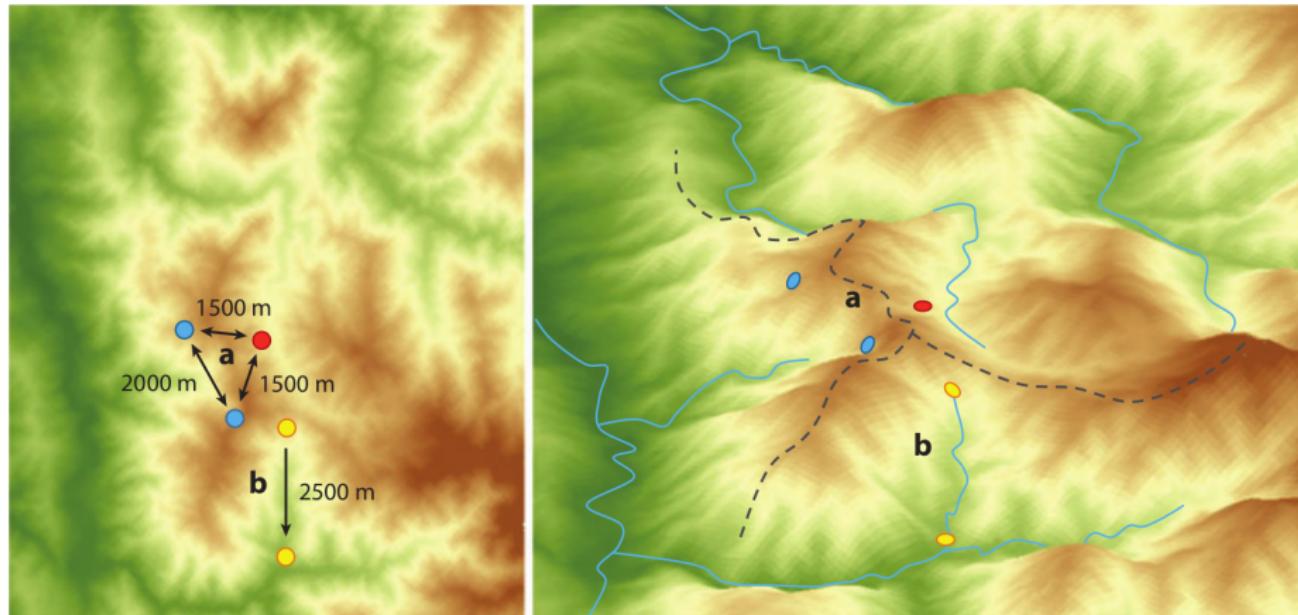
Algunos Conceptos: *Epidemiología Panorámica*

- La teledetección y su capacidad de adquirir información
- Información sobre hábitat de insectos y artrópodos

Algunos Conceptos: *Epidemiología Panorámica*

- La teledetección y su capacidad de adquirir información
- Información sobre hábitat de insectos y artrópodos
- Fuente de datos sobre la distribución espacio-temporal de enfermedades transmitidas por vectores (Pavlovsky)

Algunos Conceptos: *Epidemiología Panorámica*



Algunos Conceptos: *Epidemiología Panorámica*

- Ecología panorámica

Algunos Conceptos: *Epidemiología Panorámica*

- Ecología panorámica
- Focalidad

Algunos Conceptos: *Epidemiología Panorámica*

- Ecología panorámica
- Focalidad
 - Vectores con capacidad de transmisión de la infección

Algunos Conceptos: *Epidemiología Panorámica*

- Ecología panorámica
- Focalidad
 - Vectores con capacidad de transmisión de la infección
 - Vertebrados capaces de funcionar como reservorio de la infección

Algunos Conceptos: *Epidemiología Panorámica*

- Ecología panorámica
- Focalidad
 - Vectores con capacidad de transmisión de la infección
 - Vertebrados capaces de funcionar como reservorio de la infección
 - Huéspedes susceptibles, como humanos o animales domésticos

Algunos Conceptos: *Epidemiología Panorámica*

- Ecología panorámica
- Focalidad
 - Vectores con capacidad de transmisión de la infección
 - Vertebrados capaces de funcionar como reservorio de la infección
 - Huéspedes susceptibles, como humanos o animales domésticos

Algunos Conceptos: *Epidemiología Panorámica*

- Ecología panorámica
- Focalidad
 - Vectores con capacidad de transmisión de la infección
 - Vertebrados capaces de funcionar como reservorio de la infección
 - Huéspedes susceptibles, como humanos o animales domésticos

Epidemiología Panorámica

Algunos Conceptos: *Aprendizaje automático*

*Se dice que un programa de computadora **aprende** de experiencia E con respecto a alguna tarea T y una métrica de rendimiento M , si con la experiencia E se incrementa su rendimiento en la tarea T , medida por M .*

Tom Mitchell, 1997 [?]

Algunos Conceptos: *Aprendizaje automático (ML)*

- Enfoque empírico efectivo para *regresiones* y *clasificaciones*
- Distintos métodos:
 - Supervisados: Regresiones Lineales, SVMs, ANNs, DTRs...
 - No-supervisados: K-NNs, K-means, PCA...
 - Semi-supervisados
- Usado en muchos ámbitos:
 - Académico
 - Industrial
 - Gubernamental

Algunos Conceptos: *Métodos Supervisados*

- Aprenden a través de pares de ejemplos (X, Y_{verd})
- Conjuntos de entrenamiento y validación
- Evitar *overfitting*
- Ajuste de hiperparámetros...

Algunos Conceptos: *Parámetros y Hiperparámetros*

- Los algoritmos de ML poseen *parámetros e hiperparámetros*
 - Los hiperparámetros definen el comportamiento durante el proceso de entrenamiento
 - Los parámetros se ajustan para definir el modelo luego del entrenamiento

Algunos Conceptos: *Parámetros y Hiperparámetros*

- Los algoritmos de ML poseen *parámetros e hiperparámetros*
 - Los hiperparámetros definen el comportamiento durante el proceso de entrenamiento
 - Los parámetros se ajustan para definir el modelo luego del entrenamiento

Algunos Conceptos: *Ajuste de hiperparámetros*

Supongamos una **Ridge Regression**:

$$\hat{y}(w, x) = w_0 + w_1x_1 + \dots + w_px_p$$

$$\min_w ||Xw - y||_2^2 + \alpha ||w||_2^2$$

Algunos Conceptos: *Ajuste de hiperparámetros*

Supongamos, ahora, un pequeño **Perceptron Multicapa**:

Algunos Conceptos: *Ajuste de hiperparámetros*

Supongamos, ahora, un pequeño **Perceptrón Multicapa**:

```
MLPRegressor(activation='relu', alpha=0.0001, batch_size=  
beta_2=0.999, early_stopping=False, epsilon=1e-08,  
hidden_layer_sizes=(2, 2), learning_rate='constant',  
learning_rate_init=0.001, max_iter=200, momentum=0.0,  
nesterovs_momentum=True, power_t=0.5, random_state=None,  
shuffle=True, solver='adam', tol=0.0001, validation_fraction=0.05,  
verbose=False, warm_start=False)
```

Modelado de la población del vector de Dengue

Obtención y análisis de datos a utilizar

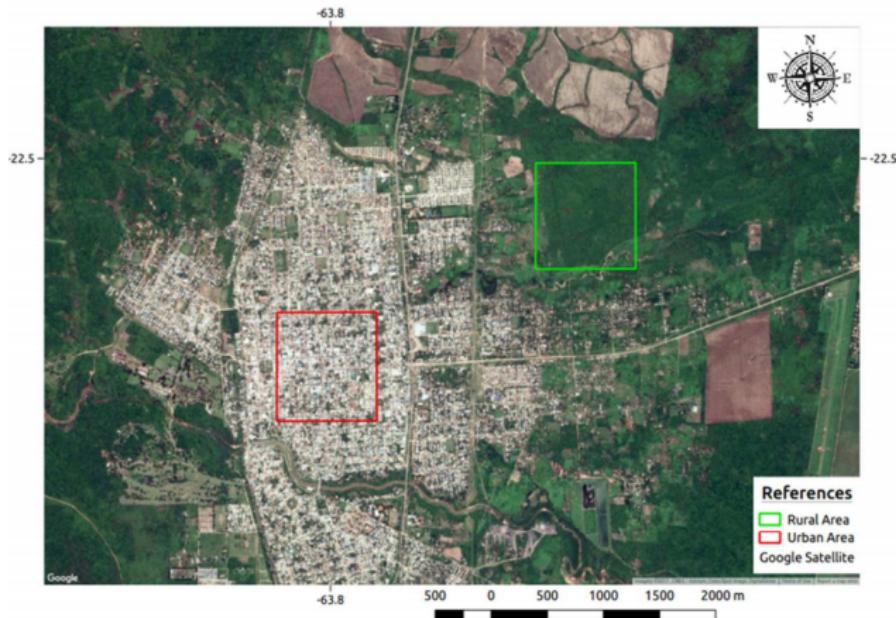


Obtención y análisis de datos a utilizar

- Oviposición: ovitrampas
- De productos satelitales a variables ambientales:
 - Propiedades de vegetación: **NDVI**
 - Humedad: **NDWI**
 - Temperatura de la superficie: **LST**
 - Precipitación: **TRMM**
- En todos los casos se contempla un *lag*

Obtención y análisis de datos a utilizar

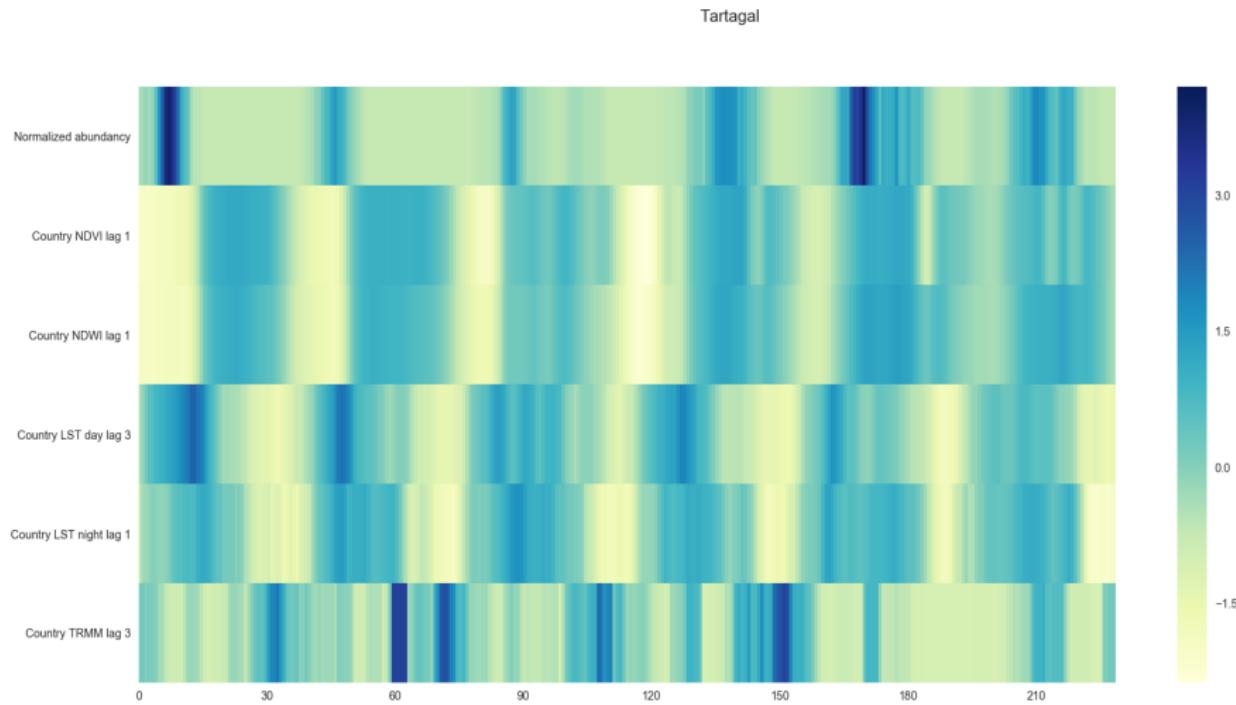
Datos de área rural y área urbana



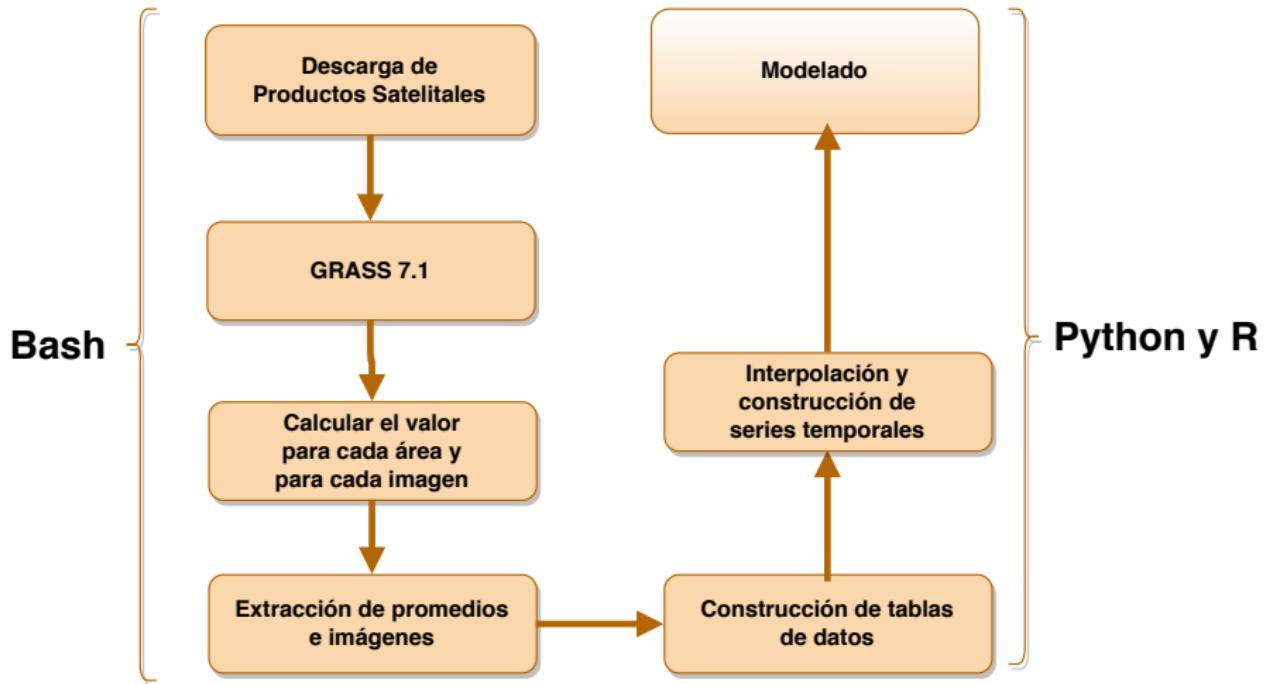
Análisis y selección de datos a utilizar

- Cinco variables elegidas:
 - NDVI rural lag 1
 - NDWI rural lag 1
 - LST rural dia lag 3
 - LST rural noche lag 1
 - TRMM lag 3
- Normalización con *z-score*

Análisis y selección de datos a utilizar



Sistema completo



Requerimientos de nuestro sistema

- Facilidad de utilización
- Herramienta de limpieza de datos
- Flexibilidad para cambiar conjuntos de datos
- Herramienta para ajustar hiperparámetros
- Generar modelos en formato utilizable en producción
- Flexibilidad para agregar nuevos algoritmos
- Herramienta de visualización de datos

Arquitectura del sistema

- Data: *scikit-learn, pandas, numpy*
- Models: *scikit-learn*
- Tuning: *Irace (Iterated Racing for Automatic Algorithm Configuration)*

Modelando poblaciones de mosquito

Todos los algoritmos de ML que se utilizaron son de la librería *scikit-learn*

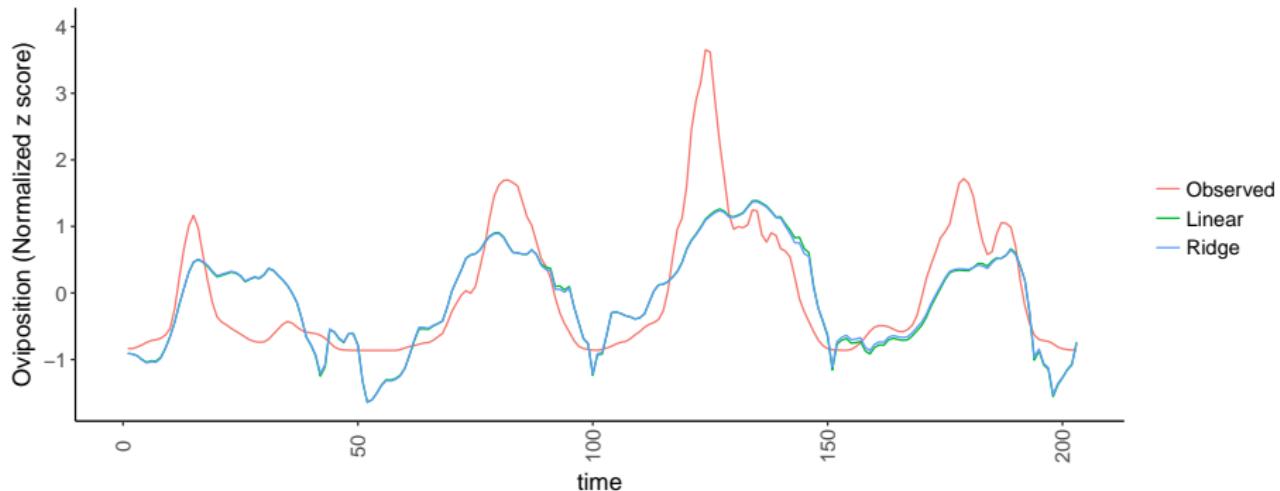
- Lineales:
 - Ridge
 - Tradicional (mínimos cuadrados)
- No-lineales:
 - Regresión de árboles de decisión (DTR)
 - Regresión de K-Vecinos más cercanos (KNNR)
 - Support Vector Regressor (SVR)
 - Perceptron multicapa (MLP)

Modelando poblaciones de mosquito

Evaluación y análisis de los modelos generados

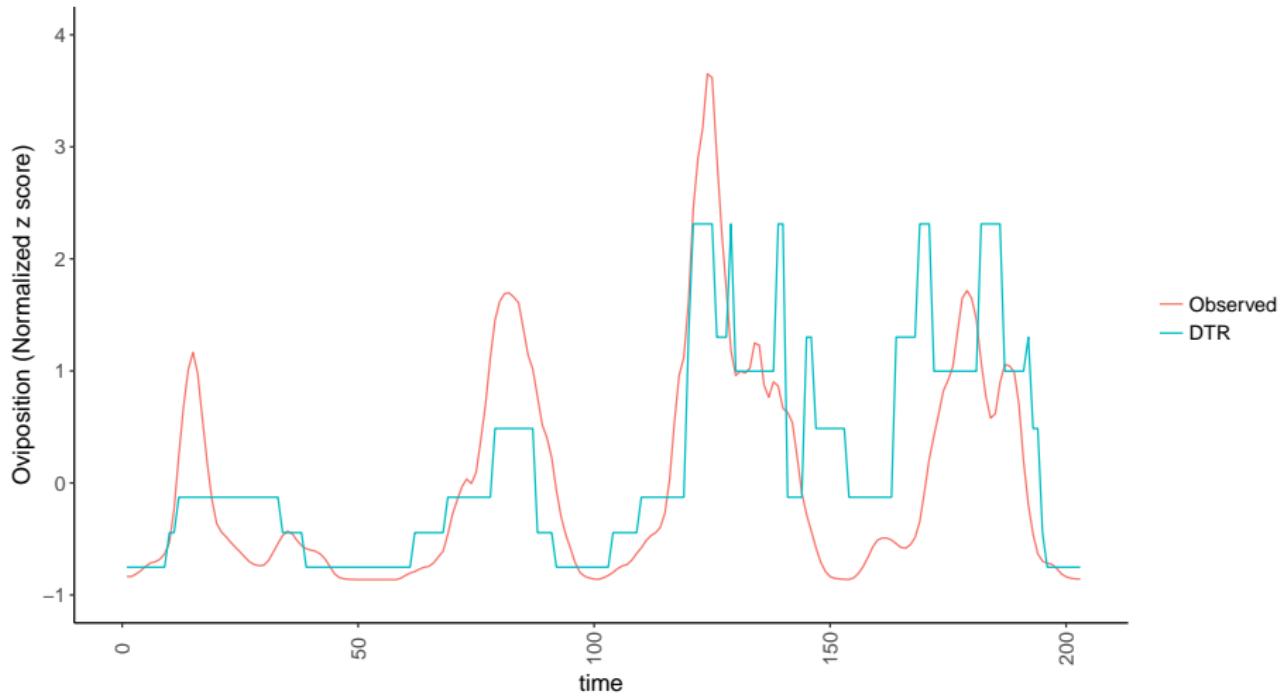
Evaluación y análisis de los modelos generados

Métodos Lineales



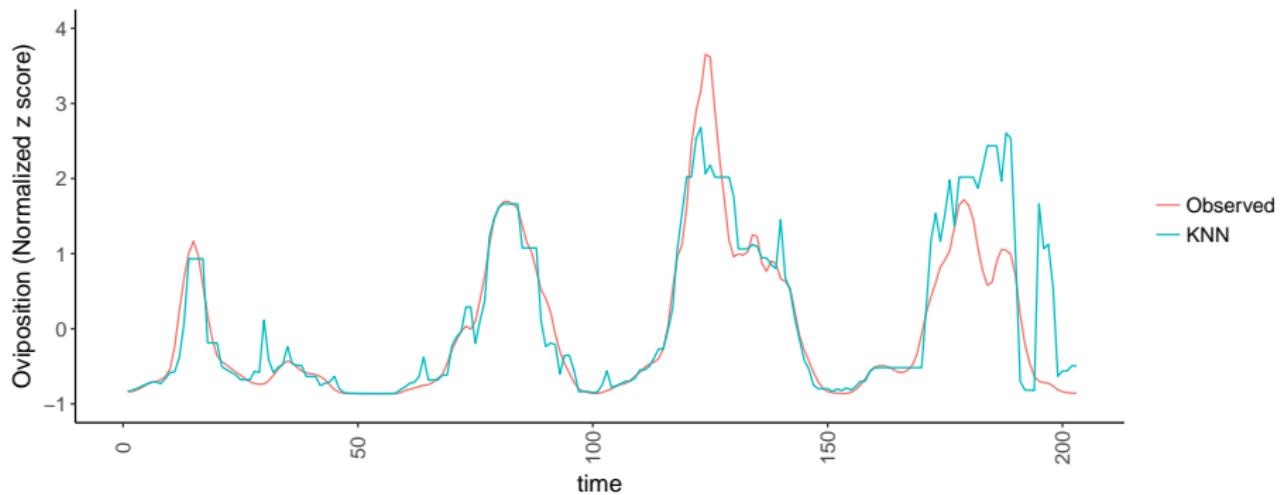
Evaluación y análisis de los modelos generados

Regresión de árbol de decisión (DTR)



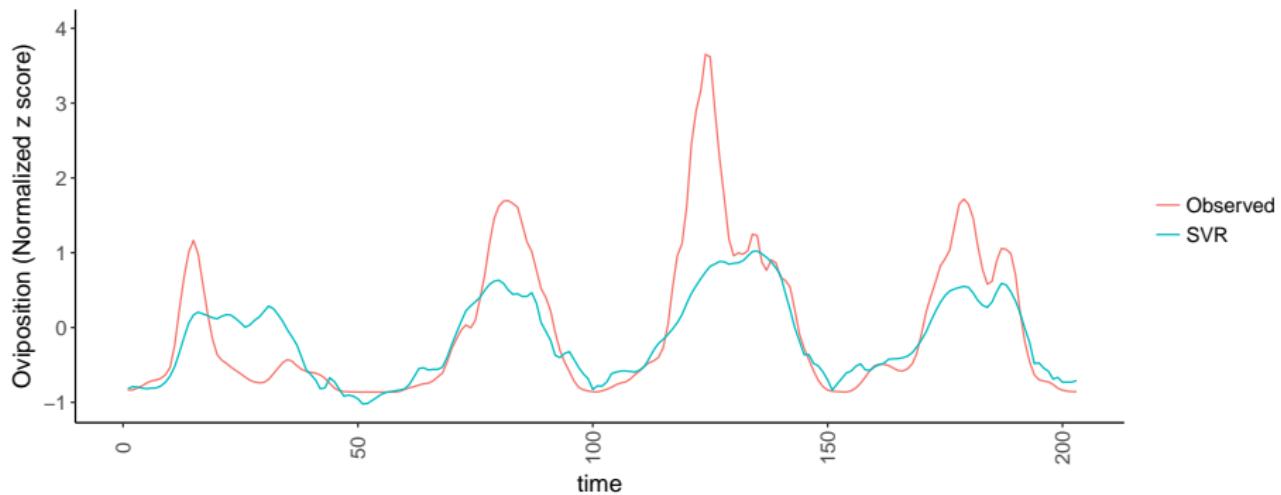
Evaluación y análisis de los modelos generados

Regresión de K-Vecinos más cercanos (KNNR)



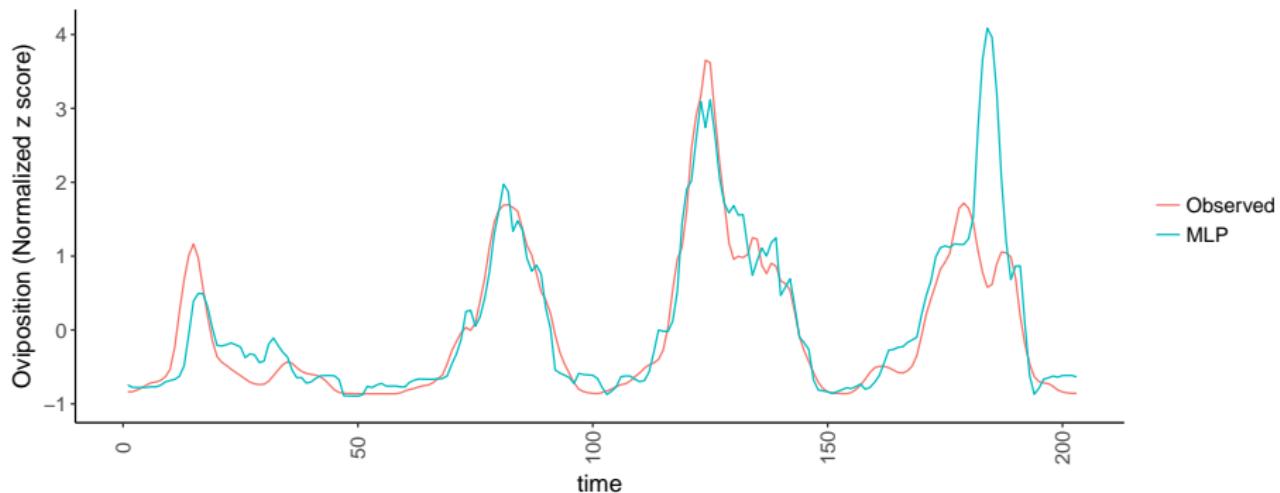
Evaluación y análisis de los modelos generados

Support Vector Regressor (SVR)



Evaluación y análisis de los modelos generados

Perceptron multicapa (MLP)



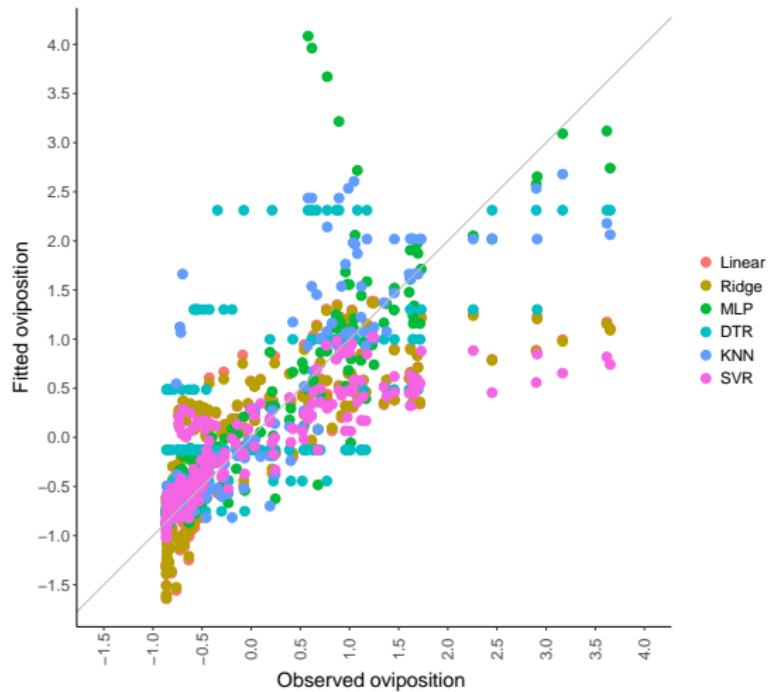
Evaluación y análisis de los modelos generados

Resumen de los datos observados y los ajustados

	Mín	$q_{1/4}$	$q_{1/2}$	Media	$q_{3/4}$	Máx
Observado	-0,863	-0,742	-0,487	0,000	0,704	3,652
Lineal	-1,641	-0,716	0,027	-0,087	0,462	1,387
Ridge	-1,638	-0,680	0,028	-0,084	0,459	1,370
MLP	-0,894	-0,677	-0,323	0,093	0,716	4,084
DTR	-0,752	-0,752	-0,128	0,138	0,998	2,312
KNNR	-0,863	-0,699	-0,501	0,099	1,033	2,679
SVR	-1,021	-0,601	-0,232	-0,147	0,309	1,023

Evaluación y análisis de los modelos generados

Scatterplot de datos observados vs datos predichos



Fin

Gracias!
Preguntas?

Referencias

- <https://github.com/juansca/WordVectors>