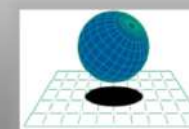


XORNADA DE USUARIOS DE R
14 Octubre, Santiago de Compostela

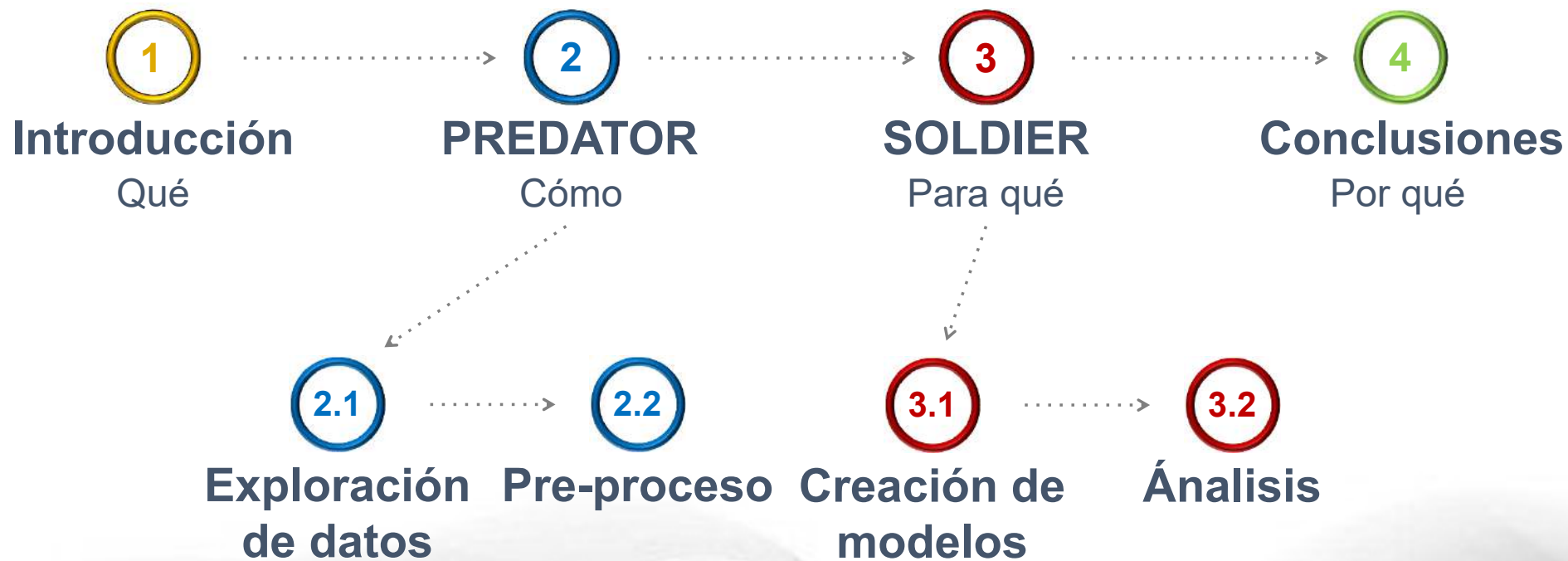


**PREPROCESAMIENTO DE DATOS DE
PRESAS ASOCIADO AL ANÁLISIS DEL
COMPORTAMIENTO MEDIANTE
APRENDIZAJE AUTOMÁTICO**



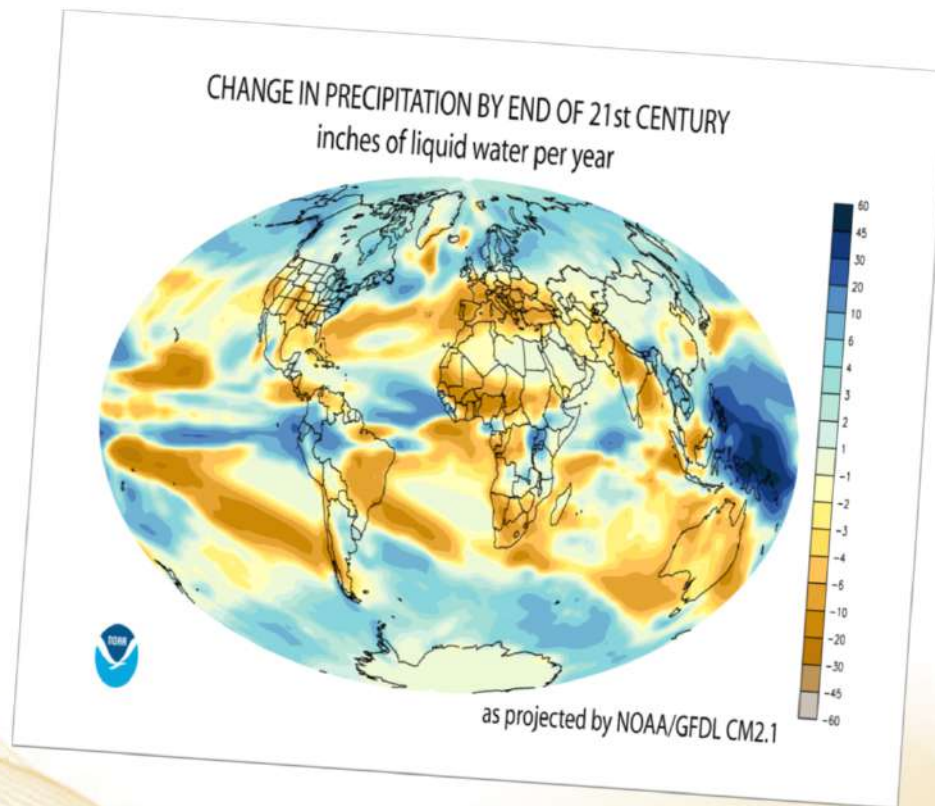
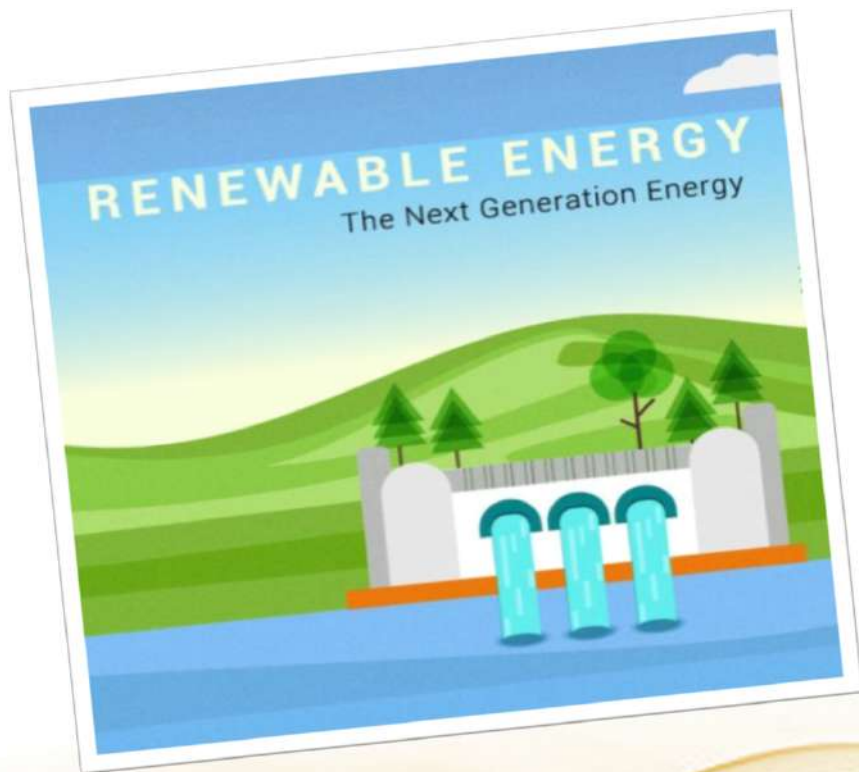
André CONDE
Fernando SALAZAR

Índice



Introducción

QUÉ



Metodología

QUÉ

Exploración

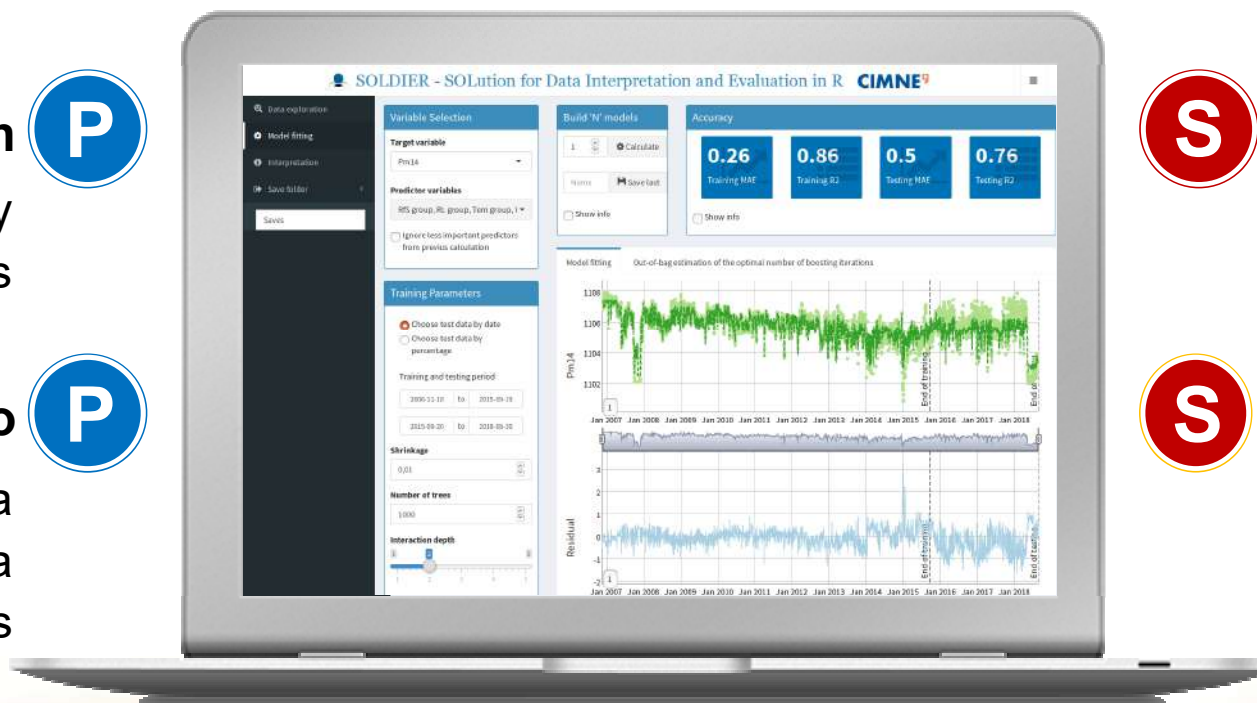


Visualización y análisis de datos

Preproceso



Generación de una base de datos para modelos predictivos



Creación de modelo

Modelos basados en machine learning



Predicción

Comprobar predicciones con datos nuevos

cimnetest.shinyapps.io/PREDATOR

cimnetest.shinyapps.io/SOLDIER

Opciones de exploración

Monitoring Data

Variables to show: Date, Month, Pm14, RfP, RL, Src Change columns to show

Show 10 entries Search: 2010

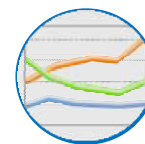
	Date	Month	Pm14	RfP	RL	Snow	SS	SumV	Temp	WF02	Year
22010	2009-03-25	03	1106.824		1116.743		0.123	90.758	-2.438	0	2009.23
28778	2010-01-01	01	1107.082		1117.145		0.12	102	-2.2	0.012	2010
28802	2010-01-02	01	1107.037		1117.352		0.12	101.92	-4.983	0.012	2010
28826	2010-01-03	01	1107.162		1117.596		0.245	105.399	-10.588	0.026	2010
28850	2010-01-04	01	1107.048		1116.894		0.154	99.22	-10.374	0	2010
28874	2010-01-05	01	1107		1117.488		0.173	102.249	-9.725	0.01	2010.01
28898	2010-01-06	01	1107.085		1117.392		0.118	102.881	-5.383	0.012	2010.01
28922	2010-01-07	01	1106.99		1116.529		0.083	96.375	-6.188	0	2010.01
28946	2010-01-08	01	1106.821		1116.507		0.11	94.738	-5.329	0	2010.01
28970	2010-01-09	01	1106.734		1116.545		0.165	94.433	-6.483	0	2010.02

Showing 1 to 10 of 367 entries (filtered from 4,304 total entries) Previous 1 2 3 4 5 ... 37 Next



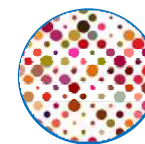
Tabla de datos

- Mostrar máximos y mínimos
- Buscar valores



Series temporales

- Dos escalas verticales
- Desplazamiento y zoom independientes



Graficos de dispersión

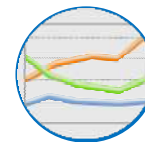
- Estándar
- Quasi-4D
- Dinámico

Opciones de exploración



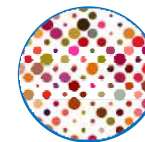
Tabla de datos

- Mostrar máximos y mínimos
- Buscar valores



Serie temporales

- Dos escalas verticales
- Desplazamiento y zoom independientes

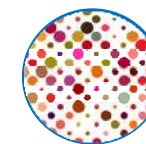
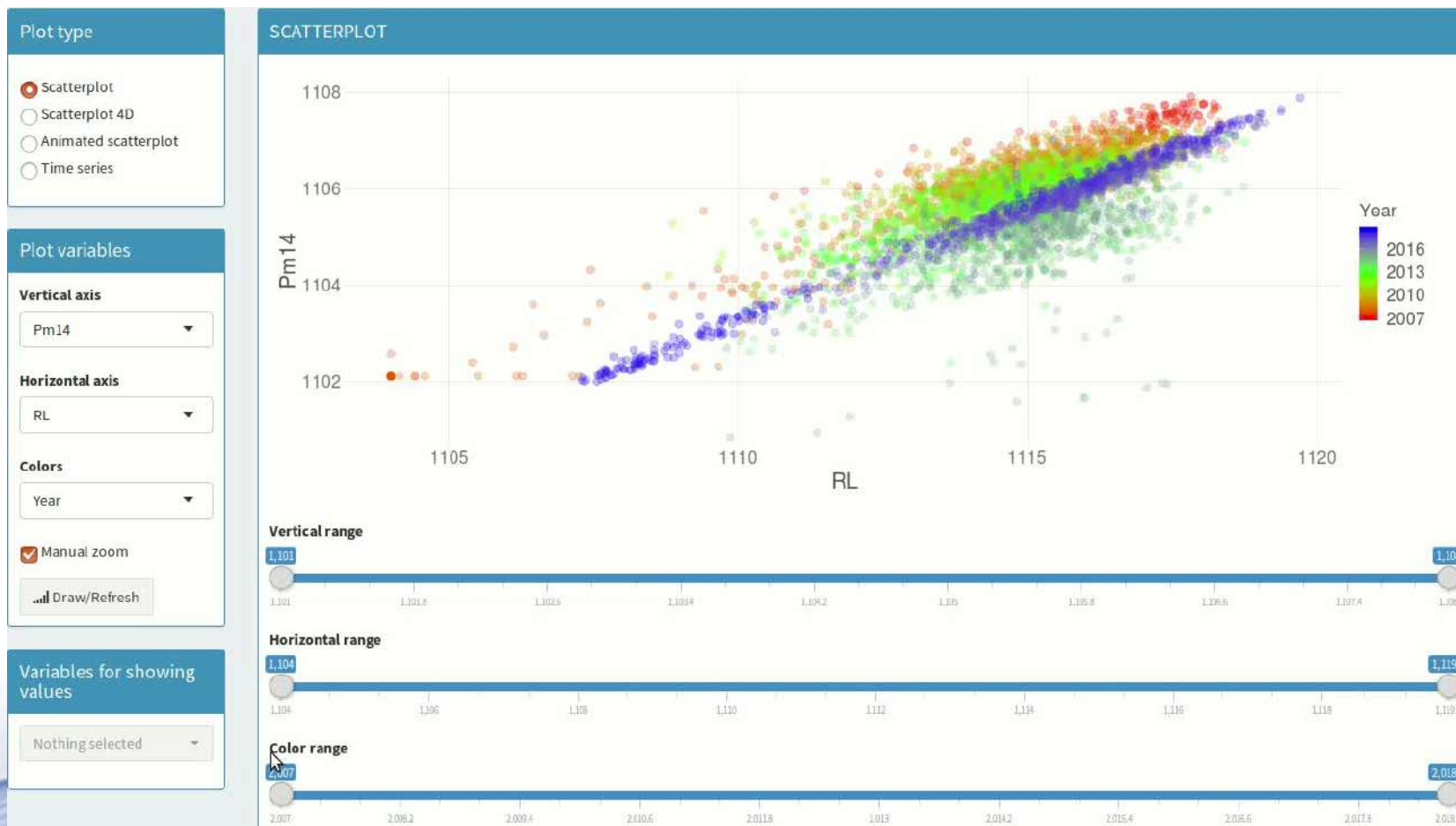


Graficos de dispersión

- Estándar
- Quasi-4D
- Dinámico

Opciones de exploración

CÓMO



Graficos de dispersión

- Estándar
- Quasi-4D
- Dinámico

Opciones de exploración

CÓMO

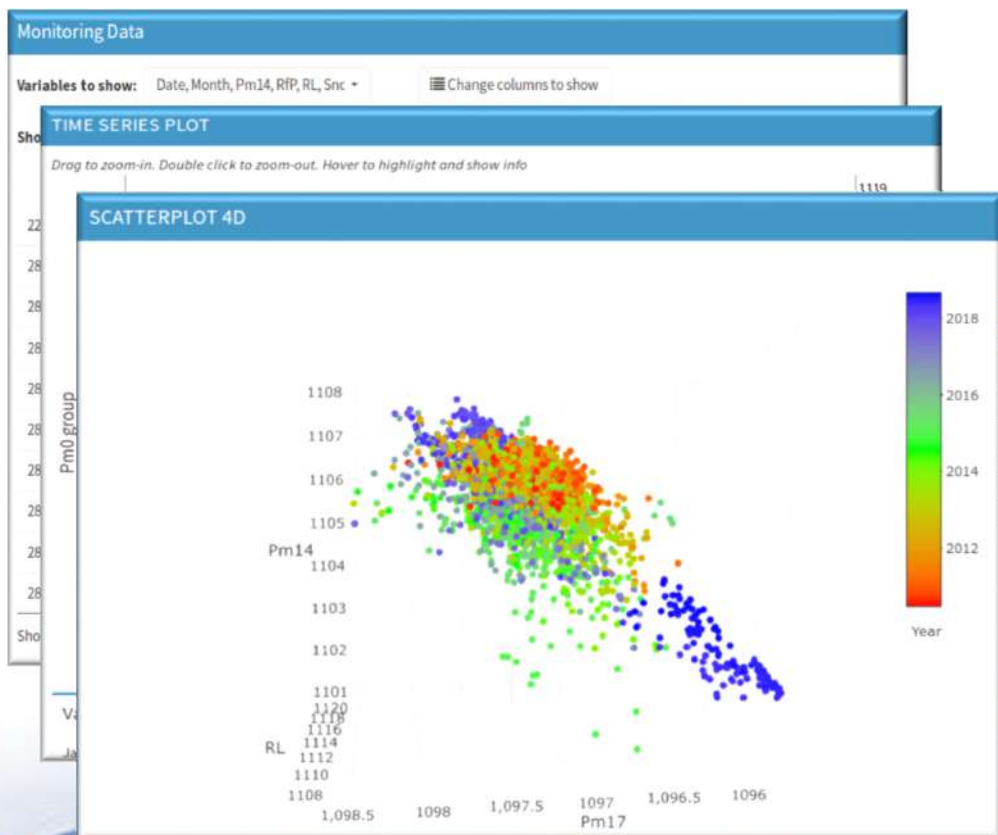
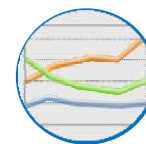


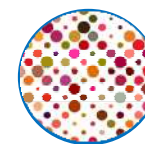
Tabla de datos

- Mostrar máximos y mínimos
 - Buscar valores



Series temporales

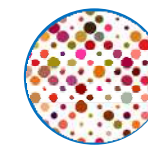
- Dos escalas verticales
 - Desplazamiento y zoom independientes



Graficos de dispersión

- Estándar
 - **Quasi-4D**
 - Dinámico

Opciones de exploración



Graficos de dispersión

- Estándar
- Quasi-4D
- Dinámico

Opciones de pre-proceso

**Datos
incompletos**

1

Eliminar variables de estudio

2

Completar valores ausentes

3

Limpieza de datos

4

Renombrar variables

5

Reducir base de datos

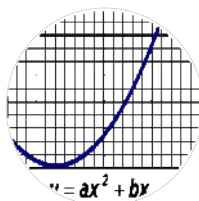
6

Crear variables derivadas

**Datos
desorganizados**

Métodos para completar

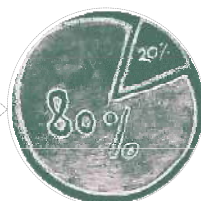
Interpolación
lineal o parabólica



Valor fijo



Estimación
(basada en
valores cercanos)



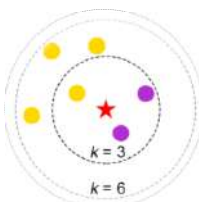
Reemplazo
(independiente de
valores cercanos)



**Media
estacional**



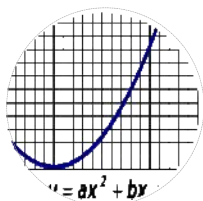
**Imputación
KNN**



La media de los valores registrados el mismo día/hora. Sólo en caso de datos con estacionalidad anual/diaria.

Limpieza de datos

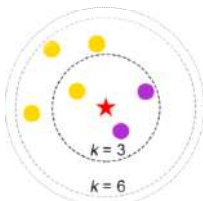
Interpolación
lineal o parabólica



Sumar una
cantidad fija



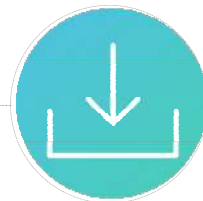
Imputación
KNN



Estimación
(basada en
valores cercanos)



Reemplazo
(independiente de
valores cercanos)



Valor fijo



Eliminar valor



Media
estacional



Permiten la compensación cuando se tiene conocimiento de un evento que afecta a los datos y se desea que este evento no influya en los resultados

Limpieza de datos

CÓMO

Interpolación
 lineal o parabólica

Sumar una
 cantidad fija

Imputación
 KNN

Choose variable

Reservoir_dm_value..ft. ▾

Manual zoom

Modify values Show info

Changes on selected area:

Modify existing values

Fill missing values

Methodology:

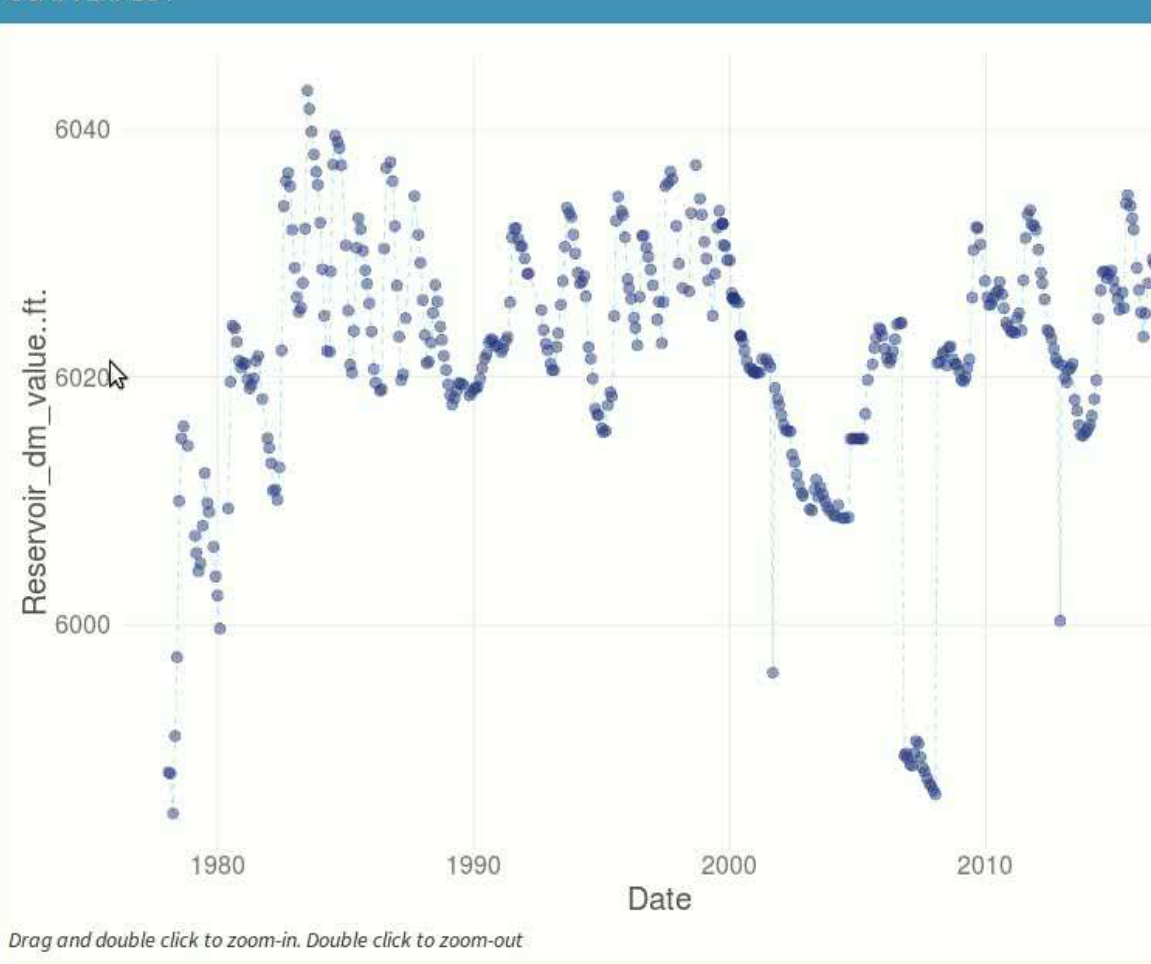
Lineal interpolation ▾

Change values

Undo changes on selected points

Save modifications for this variable

SCATTERPLOT



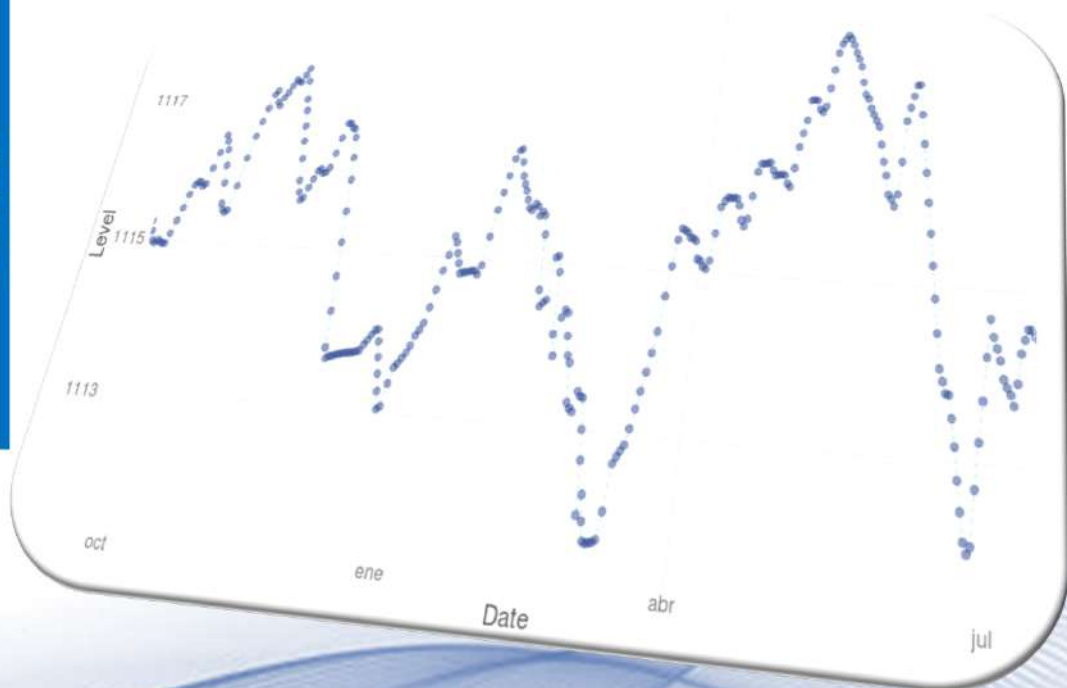
valor fijo

eliminar valor

Media
 sustitucional

Drag and double click to zoom-in. Double click to zoom-out

Reducir base de datos



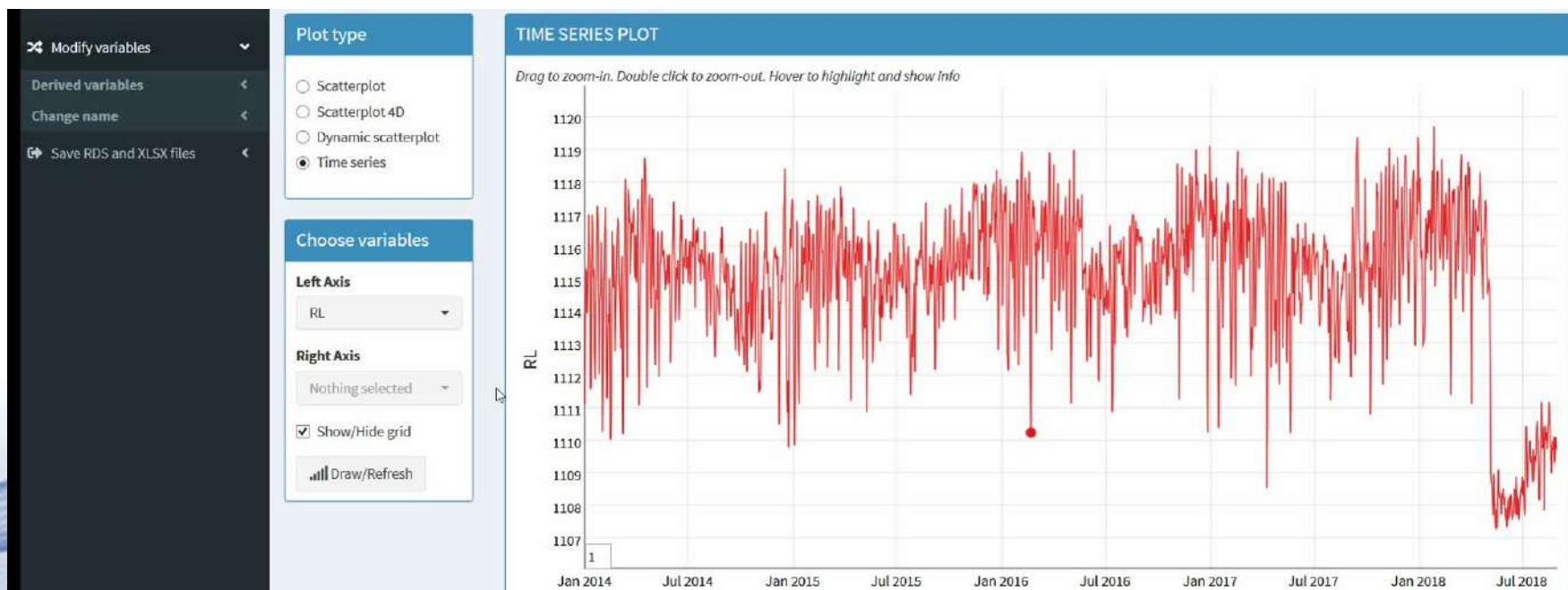
Información irrelevante o redundante puede conducir a modelos predictivos menos eficientes.

PREDATOR permite diferentes maneras de reducir los datos a un solo valor por día, semana, quincena o mes:

- Valor máximo o mínimo para cada período
- Media o suma de los valores de cada período
- Valor registrado a una hora determinada del día

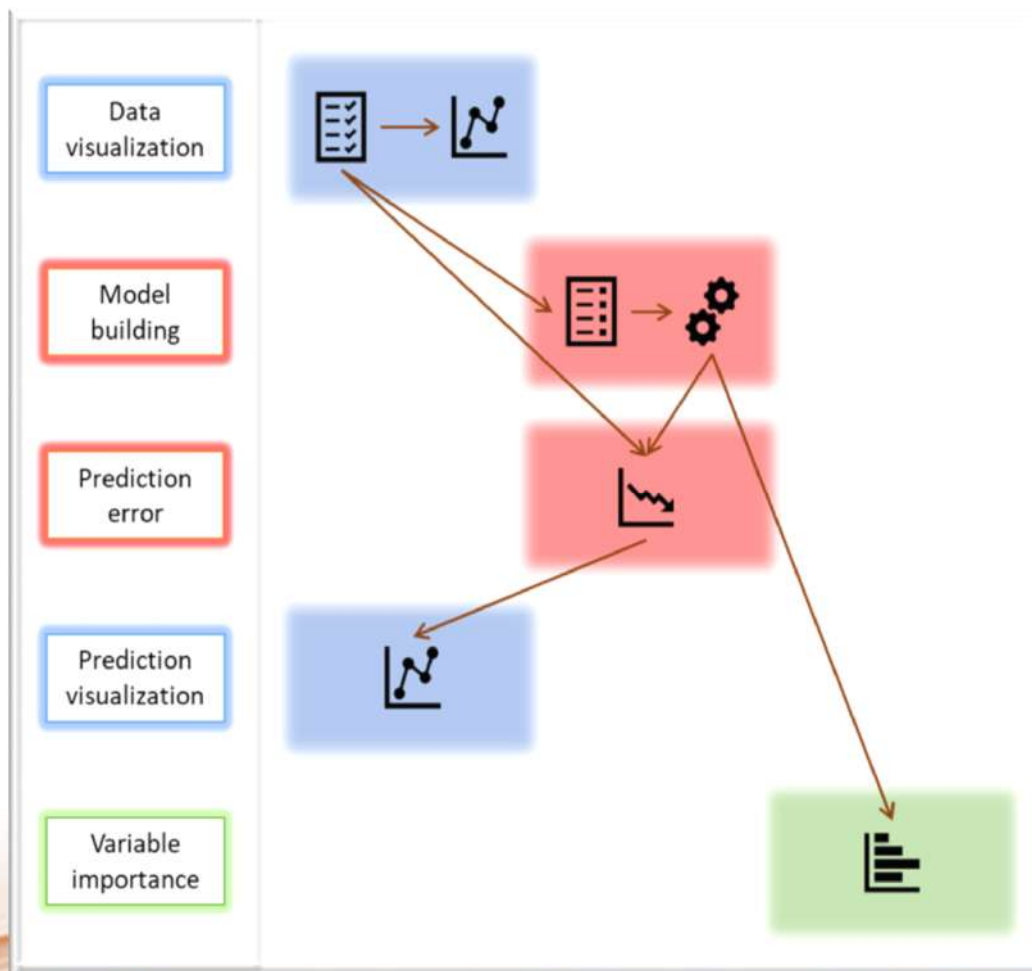
VARIABLES DERIVADAS

Se pueden generar nuevas variables: calculando **sumas acumuladas**, **combinaciones** de varias variables o los **incrementos** de una existente. Si se dispone de variable temporal se pueden crear los **meses** o los **años**. Del mismo modo se pueden calcular **medias móviles** (línea azul) de variables existentes (línea roja).



Análisis de registros

PARA QUÉ



Complete data



Selected data



Error analysis



Visual analysis



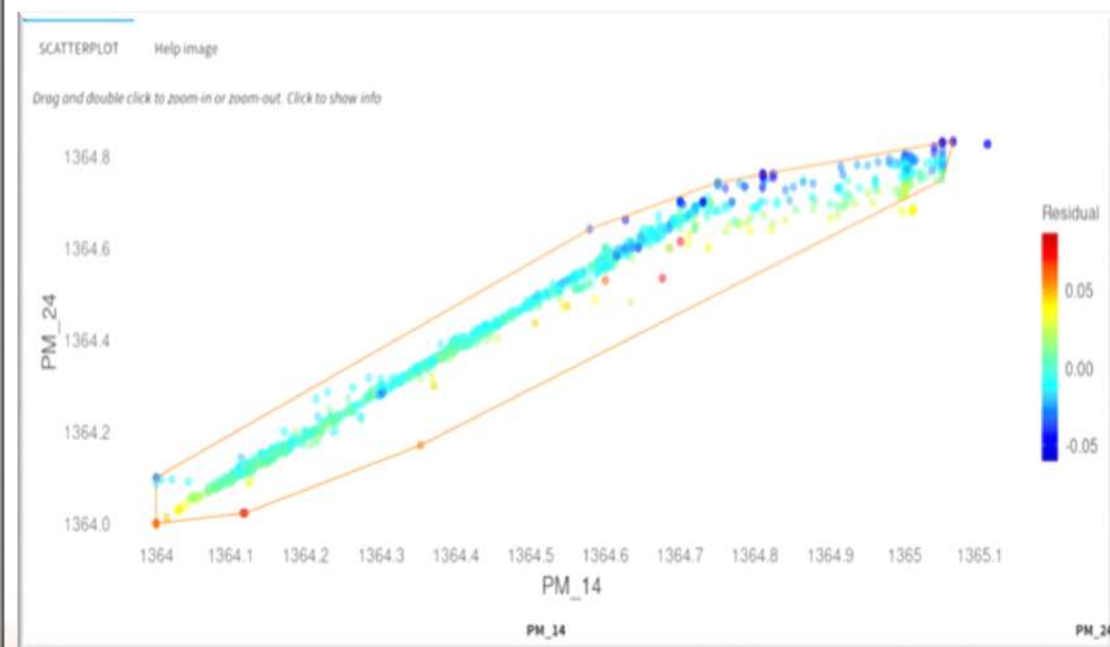
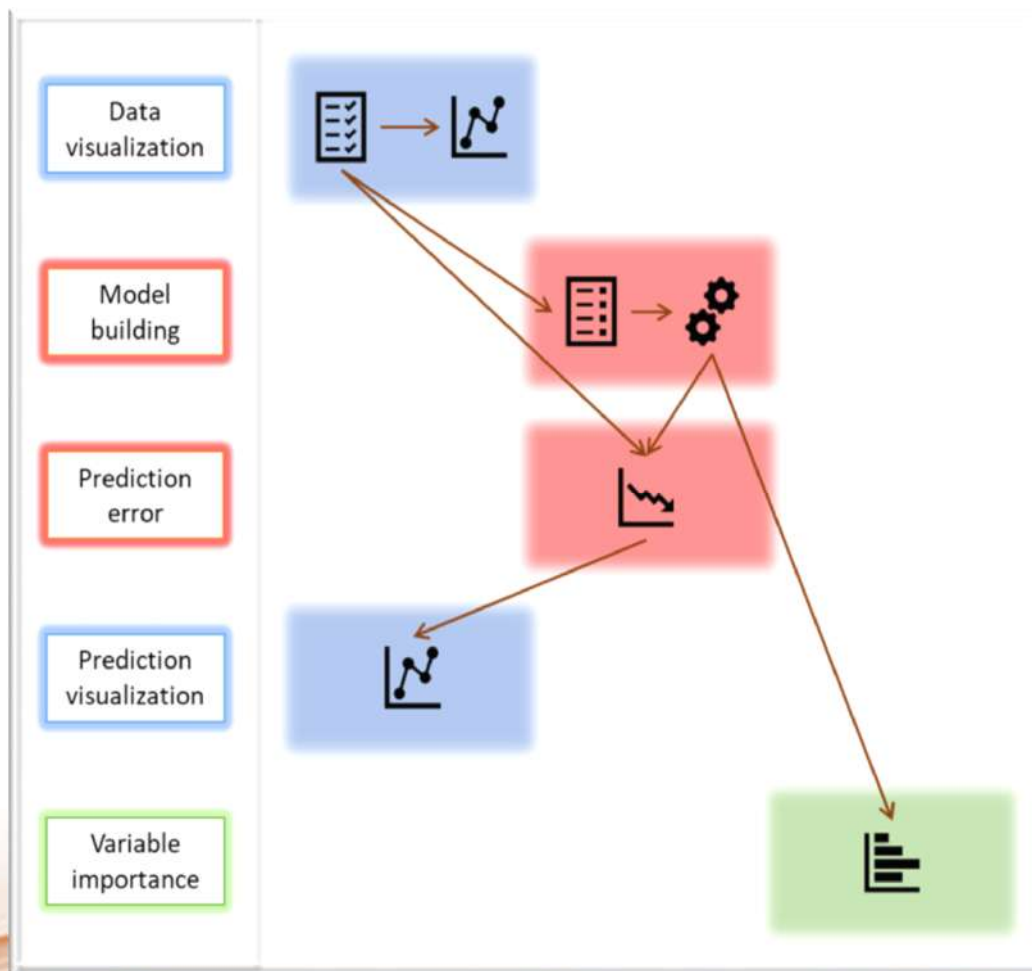
Create model



Importance analysis

Análisis de registros

PARA QUÉ



Aprendizaje automático

PARA QUÉ



Selecciona automáticamente las variables más relevantes. Por lo tanto, las variables de entrada correlacionadas se pueden considerar sin pérdida de precisión

Creación de modelos

Variable Selection

Target variable

Pm14 ▼

Predictor variables

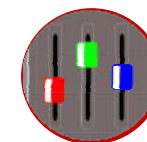
RfP group, RfS group, RfT group, I ▼

Ignore less important predictors from previous calculation



Variables

- Variable objetivo
- Cualquier cantidad de variables
- Cantidad de modelos



Modelo ML

- Períodos de entrenamiento y comprobación
- Parametros del modelo

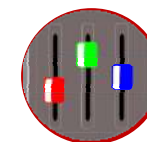
Creación de modelos

PARA QUÉ



Variables

- Variable objetivo
- Cualquier cantidad de variables
- Cantidad de modelos



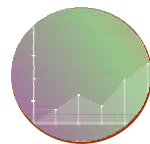
Modelo ML

- Períodos de entrenamiento y comprobación
- Parametros del modelo

Fiabilidad del modelo

PARA QUÉ

New data	
Flow	2,2275
Piezometer	1105,879
SumV	78,8265
Susp_Solid	0,3075
Temperature	5,031
Well	1102,241
Year	2017,815



Predicción

- Base de datos
- Valores seleccionados



Error

- Coeficiente de determinación R²
- Error absoluto medio MAE

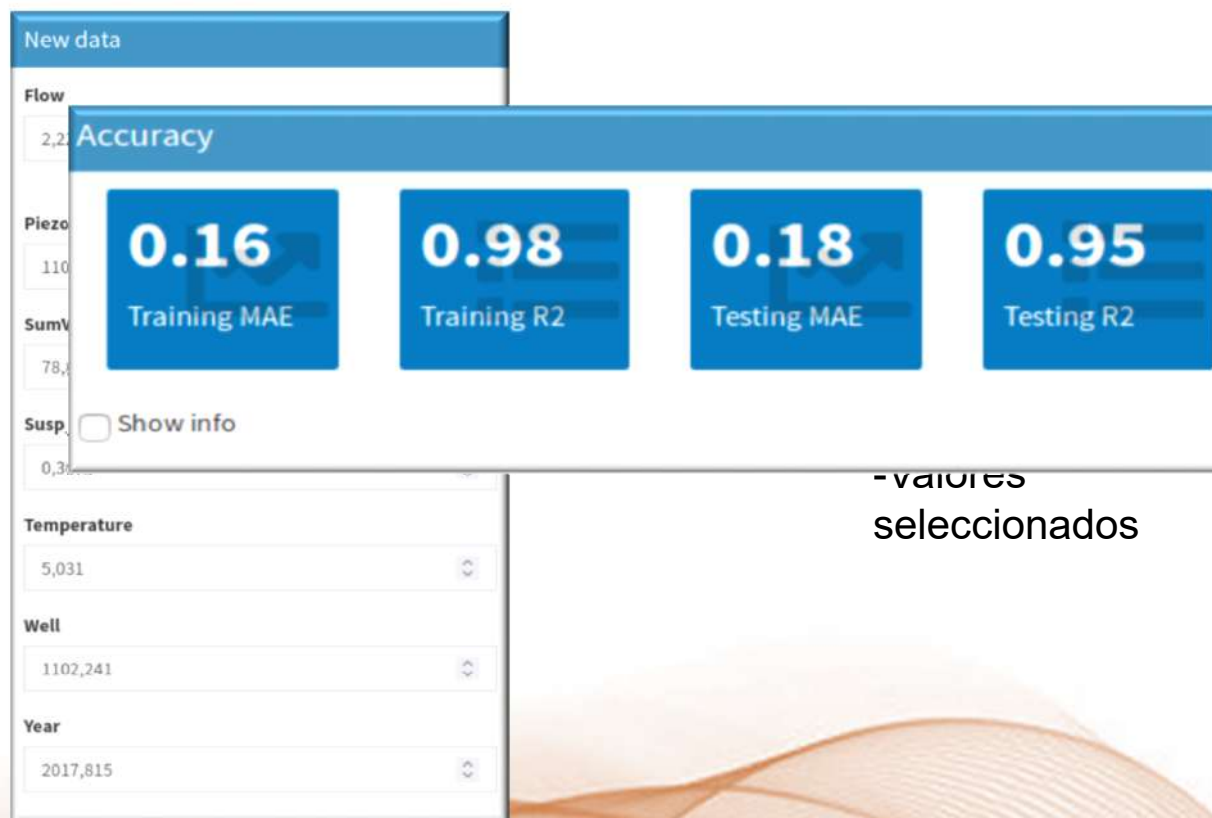


Comparación

- Datos vs predicción
- Residuos datos/predicción

Fiabilidad del modelo

PARA QUÉ



-valores
seleccionados



Error

- Coeficiente de determinación R2
- Error absoluto medio MAE



Comparación

- Datos vs predicción
- Residuos datos/predicción

Fiabilidad del modelo

PARA QUÉ



Error

- Coeficiente de determinación R²
- Error absoluto medio MAE



Comparación

- Datos vs predicción
- Residuos datos/predicción

Fiabilidad del modelo

PARA QUÉ



Error

- Coeficiente de determinación R²
- Error absoluto medio MAE

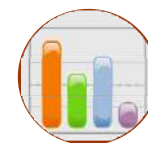
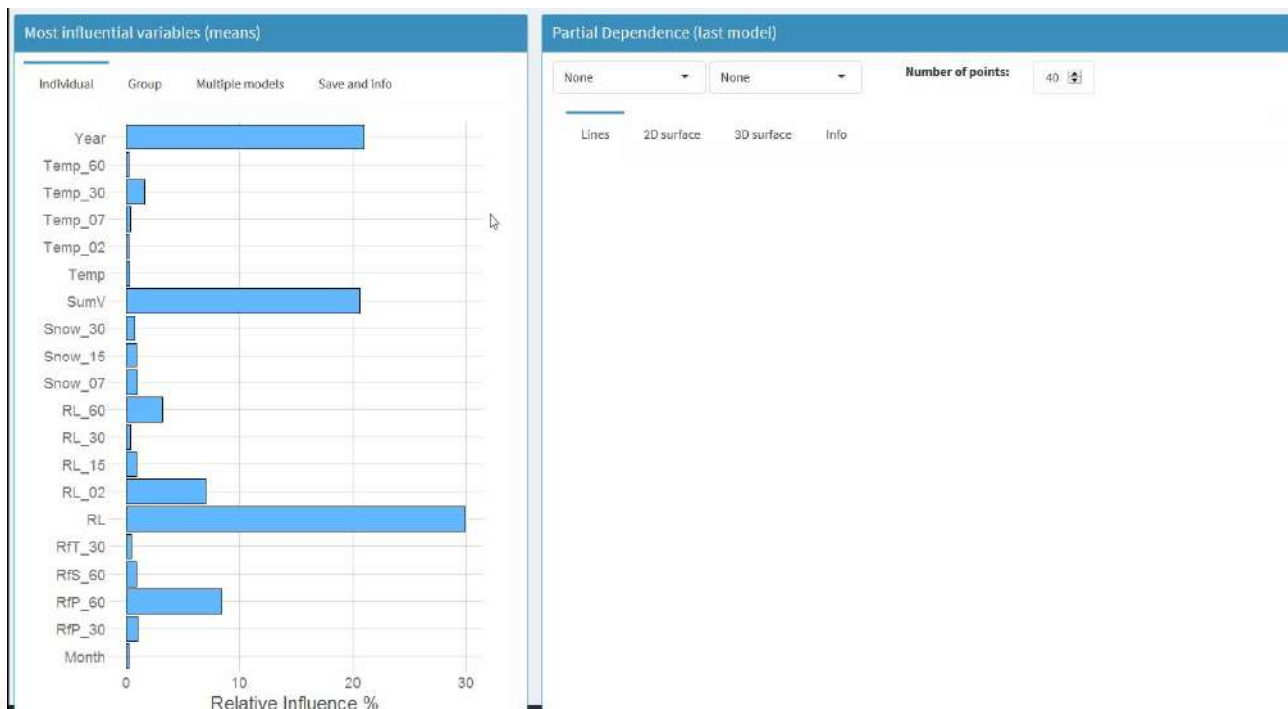


Comparación

- Datos vs predicción
- Residuos datos/predicción

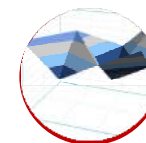
Análisis de resultados

PARA QUÉ



Influencia de variables

- Influencias individuales
- Influencias por grupos
- Múltiples modelos



Dependencia parcial

- PDP 1D
- PDP 2D
- PDP 3D**

Utilidades

01

Identificación de cambios en el comportamiento pasado del sistema

02

Detección de desviaciones del funcionamiento normal en tiempo real

03

Determinación de la asociación entre variables de entrada y la respuesta

04

Predicción de la respuesta de la presa a un conjunto de acciones

05

Localización de las zonas de mayor sensibilidad

Ejemplo de aplicación

Primero usamos los datos brutos de las variables externas de una presa para construir un modelo predictivo para el nivel en un piezómetro.

Posteriormente construimos el mismo modelo utilizando PREDATOR para preprocesar los datos, y con ello mejoran tanto los errores en los datos de entrenamiento como los errores en los datos de prueba.

The screenshot displays the SOLDIER web application interface, titled "SOLDIER - SOLUTION for Data Interpretation and Evaluation in R" with the CIMNE logo. The interface is divided into several sections:

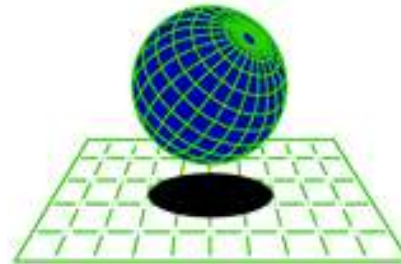
- Variables Selection:** Features a "Target variable" dropdown menu set to "Pm14" and a "Predictor variables" dropdown menu set to "Nothing selected". There is also a checkbox for "Ignore less important predictors from previous calculation".
- Build 'N' models:** Includes a numeric input field set to "1", a "Calculate" button, a "Name" input field, and a "Show info" checkbox.
- Training Parameters:** Contains radio buttons for "Choose test data by date" (selected) and "Choose test data by percentage". It also has two date range inputs for "Training and testing period": "2014-01-01 to 2017-06-30" and "2017-07-01 to 2018-08-30". A "Shrinkage" input field is set to "0,01".
- Accuracy:** Displays four blue buttons for "Training MAE", "Training R2", "Testing MAE", and "Testing R2". A "Show info" checkbox is located below these buttons.
- Model fitting:** A section titled "Out-of-bag estimation of the optimal number of boosting iterations" with a "Show/Refresh results" button.

Autores

André Conde Vázquez - aconde@cimne.upc.edu

Fernando Salazar González - fsalazar@cimne.upc.edu

International Center for Numerical Methods in Engineering, Spain



CIMNE[®]

cimnetest.shinyapps.io/PREDATOR

cimnetest.shinyapps.io/SOLDIER

