

Predicción e interpolación dinámica de los niveles de contaminación en base al tráfico y la dirección del viento

Lidia Contreras
César Ferri

liconoc@upv.es
cferri@dsic.upv.es

@liconoc
@ceferra



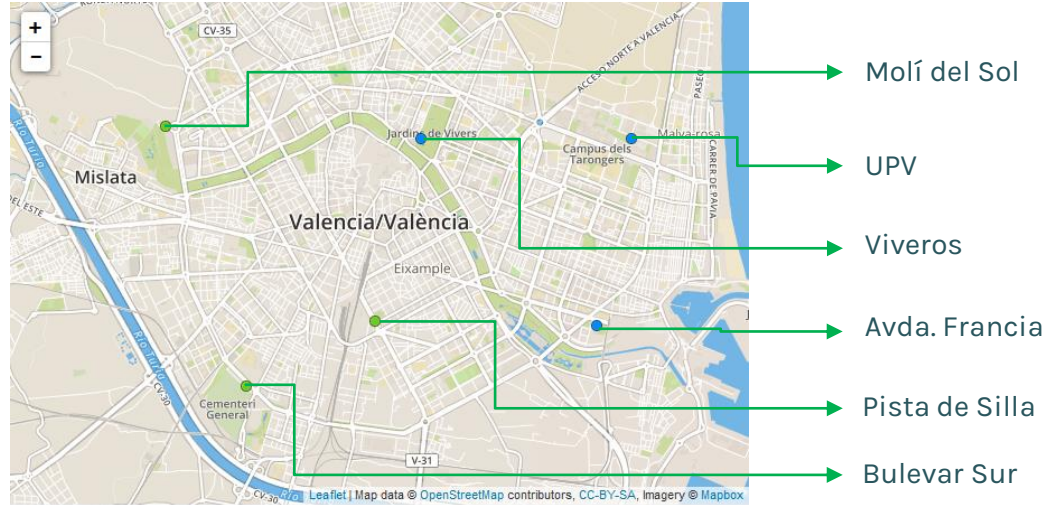
UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



DSIC
DEPARTAMENT DE SISTEMES
INFORMÀTICS I COMPUTACIÓ



Introducción



- 6 estaciones de medición de contaminación
 - 3 horas para validación de datos
 - No existen datos públicos en tiempo real
- Objetivos:
- Predecir el nivel de contaminación en las seis estaciones mediante datos de tráfico y datos meteorológicos en tiempo real.
 - Extrapolar el resultado a toda la ciudad

Selección de contaminantes

	Año	CO	NO	NO2	PM2.5	PM10	SO2	O3
Bulevar	2013	25.38	72.37	72.37			82.25	97.26
	2014	89.21	89.21	89.21			89.49	98.69
	2015	8.46	86.46	86.46			90.23	97.66
	TOTAL	41.02	82.68	82.68	0	0	87.32	97.87
Francia	2013	90.7	79.95	79.95	93.43	93.37	84.89	97.43
	2014	81.95	81.11	81.11	92.6	92.6	76.63	92.6
	2015	90.13	84.19	84.19	77.64	77.64	85.11	97.04
	TOTAL	87.59	81.75	81.75	87.89	87.87	82.21	95.69
Moli	2013	48.13	79.59	79.59	72.08	72.08	94.16	96.61
	2014	88.25	89.18	88.84	96.63	96.63	65.66	94.17
	2015	74.09	86.39	86.27	84.91	84.91	96.1	95.76
	TOTAL	70.16	85.05	84.90	84.54	84.54	85.31	95.51
Pista	2013	73.56	69.43	69.43		99.79	71.51	99.27
	2014	96.28	99.69	99.69	99.34	99.34	89.83	97.23
	2015	86.73	88.97	88.97	64.23	97.49	86.28	89.68
	TOTAL	85.52	86.03	86.03	54.52	98.87	82.54	95.39
UPV	2013		84.26	84.26	97.31	97.31	68.2	82.93
	2014		96.68	96.68	97.93	97.93	88.65	93.42
	2015		92.82	92.82	97.9	97.9	83.36	91.07
	TOTAL	0	91.25	91.25	97.71	97.71	80.07	89.14
Viveros	2013	88.39	63.25	63.25			74.24	85.53
	2014		92.12	92.12			91.02	97.19
	2015	29.46	81.9	81.9			83.74	90.41
	TOTAL	39.28	79.09	79.09	0	0	83.00	91.04

Porcentaje de datos completos en los históricos

Fuentes de datos y parámetros

Nivel de contaminantes:

- Fuente: Generalitat Valenciana
- Parámetros: NO, NO₂, O₃, SO₂

Condiciones meteorológicas:

- Fuente: Generalitat Valenciana
- Parámetros: Temperatura, humedad relativa, presión, velocidad del viento, precipitaciones.

Intensidad del tráfico:

- Fuente: Ayuntamiento de Valencia (1231 espiras electromagnéticas).
- Parámetros (vehículos/hora): Nivel de tráfico alrededor de la estación.

Características temporales:

- Año, mes, día del mes, día de la semana, hora.



Agregación de datos

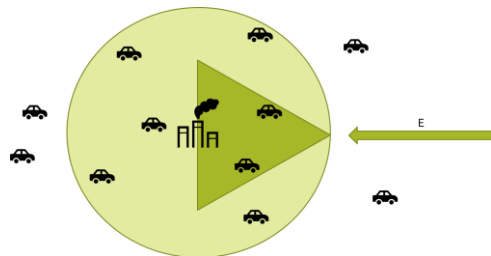
Para cada estación de medición de contaminación:

1. Juntar datos de contaminación + datos meteorológicos (por fecha)

```
data_contaminacion$day<-as.numeric(format(as.Date(data_contaminacion$FECHA,"%d/%m/%Y"),"%d"))
data_contaminacion$month<-as.numeric(format(as.Date(data_contaminacion$FECHA,"%d/%m/%Y"),"%m"))
data_contaminacion$year<-as.numeric(format(as.Date(data_contaminacion$FECHA,"%d/%m/%Y"),"%Y"))

datos_unidos<-merge(data_clima,data_contaminacion,by=c("day","month","year","HORA"))
```

2. Calcular el tráfico asociado a la estación (radio = 1km) cada hora, según la dirección del viento.



Librería: geosphere

Distancia de la estación al punto de tráfico (LAT, LON):

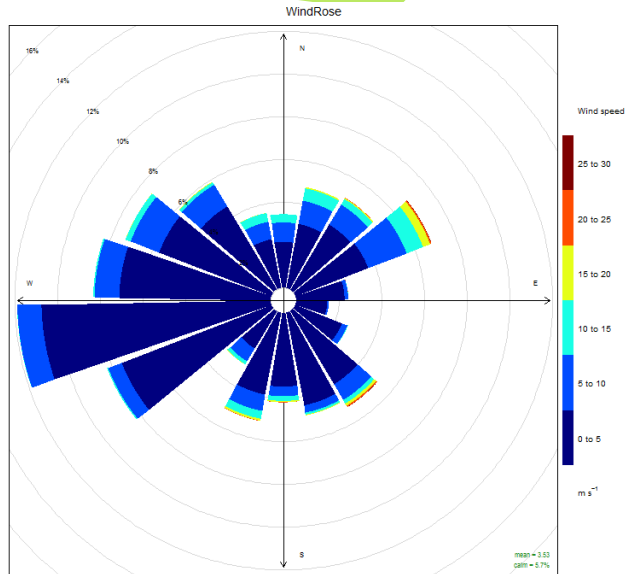
`distHaversine(Estación, Punto)` → metros

Dirección del punto de tráfico respecto a la estación (LAT, LON):

`bearingRhumb(Punto, Estación)` → grados

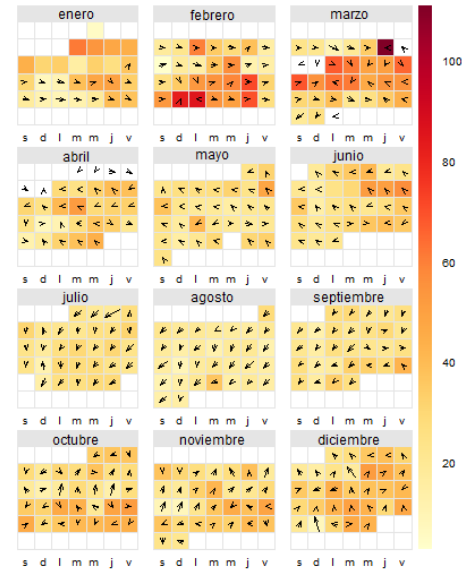
3. Generar un archivo csv con los datos de los tres años.

Estudio de los datos



```
windRose(dat, breaks=c(0,5,10,15,20,25,30),  
angle=20, cols="jet", paddle=FALSE, offset=5,  
key.header="Wind speed", key.footer="m/s",  
key.position="right", main="WindRose")
```

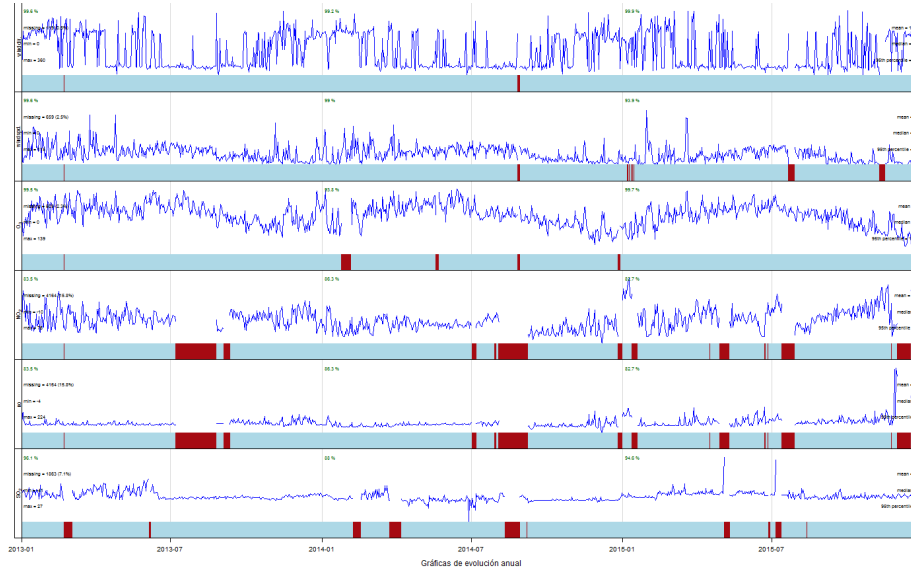
Dirección del viento en 2014 (+NO₂ en Bulevar)



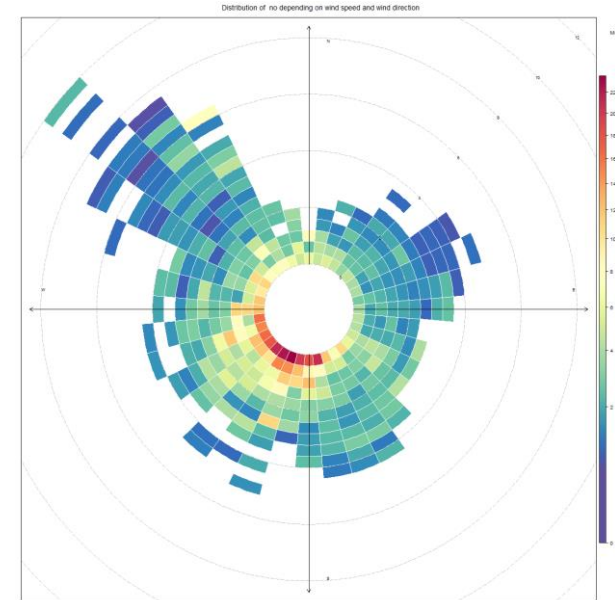
```
calendarPlot(dat, pollutant =  
"No2", year = 2014, annotate =  
"ws", main="Dirección del viento  
en 2014 (+NO2 en Bulevar)")
```

Librería: openair

Estudio de los datos

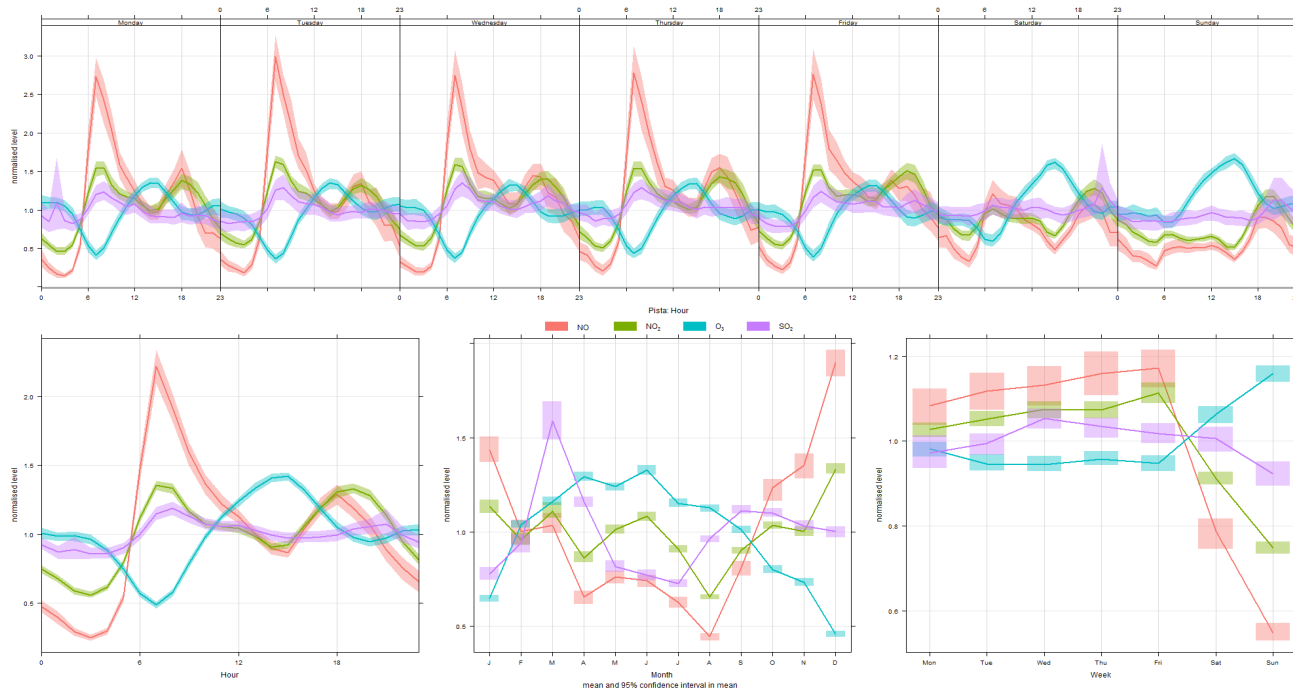


```
summaryPlot(dat, percentile=0.98, na.len=12,
col.trend="blue", xlab=c("Gráficas de evolución
anual", "Histogramas"), main="Gráfica resumen de
parámetros")
```



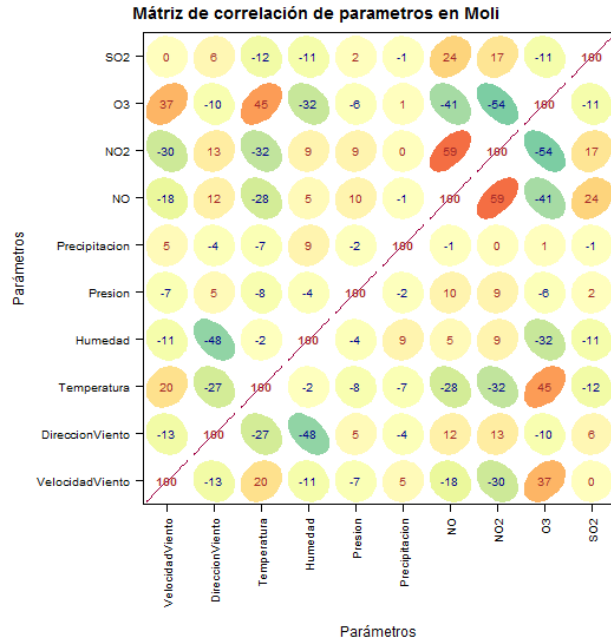
```
polarFreq(dat, pollutant="NO",
statistic="mean", ws.int=0.4,
ws.upper=8, min.bin=5, grid.line=2,
border.col="white", main="Distribution
of no depending on wind speed and wind
direction",key.header="Mean",
key.footer="")
```

Estudio de los datos



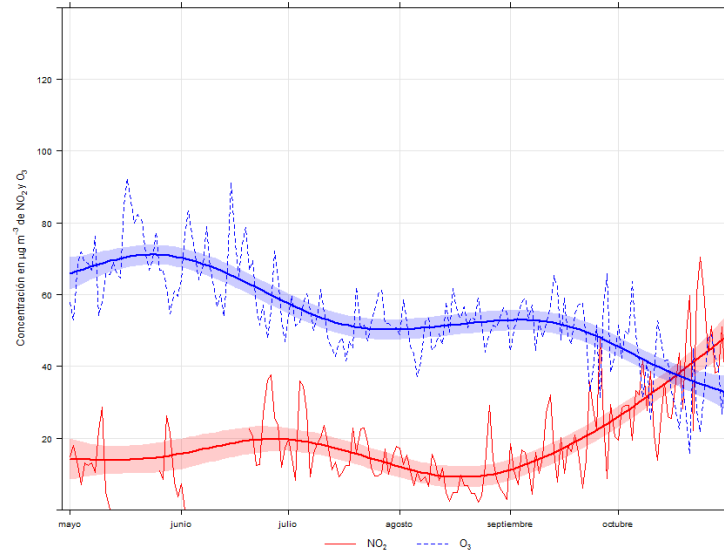
```
timeVariation(dat, pollutant=objs2[1:4], normalise = TRUE, difference=FALSE,  
ylab="Concentration", xlab=c(paste(estaciones[i],": Hour",sep = ""), "Hour",  
"Month", "Week"))
```


Estudio de los datos



```
corPlot (dat, type="season", layout=c(2,2),
cluster=FALSE, main="MATRIZ DE CORRELACIÓN
DE PARÁMETROS", xlab="Parámetros",
ylab="Parámetros", auto.text=FALSE,
text.col=c("darkblue", "brown"))
```

EVOLUCIÓN MEDIAS DIARIAS NO₂ Y O₃ - Verano 2014



```
timePlot (selectByDate(dat, year=2014,
month=c(5,6,7,8,9,10)), pollutant=c("NO2","O3"),
group=TRUE, avg.time="day", cols=c("red", "blue"),
smooth=TRUE, date.breaks=6, date.format="%B",
ylab="Concentración en ug/m3 de NO2 y O3",
main="EVOLUCIÓN MEDIAS DIARIAS NO2 Y O3 - Verano
2014", ylim=c(0,140))
```

Predicción

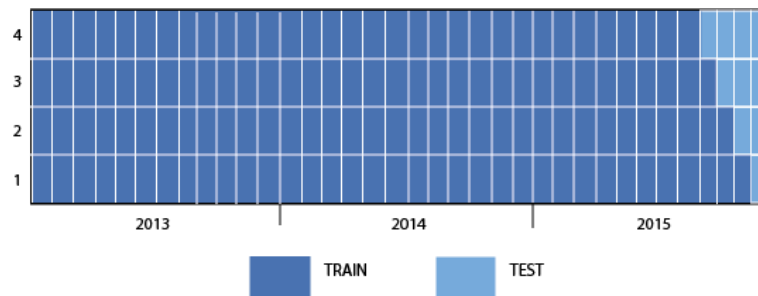
Modelos de regresión:

- Linear Regression (lr)
- Quantile Regression (qr) con lasso method
- K nearest neighbours (IBKreg) con k=10
- Decision tree for regression (M5P)
- Random Forest (RF)
- Tres modelos básicos:
 - TrainMean: Siempre predice la media de Train
 - TestMean: Siempre predice la media de Test
 - X3H: Valor del contaminante 3h antes

Librerías: Rweka, quantreg, randomForest

Data Splitting:

- Test: 1, 2, 3 y 4 meses
- Train: Dataset-Test



Librería: caret

```
#24h al día * 30 días = 720h al mes  
#mesestest va de 1 a 4  
numtest=mesestest*720  
timeSlices<-createTimeSlices(1:nrow(dat),  
initialWindow = nrow(dat)-numtest, horizon =  
numtest, fixedWindow = FALSE, skip = 720)
```

Resultados

	<i>TrainMean</i>	<i>TestMean</i>	<i>X3h</i>	<i>LinearRegression</i>	<i>qr</i>	<i>IBkreg</i>	<i>M5p</i>	<i>RF</i>
<i>NO</i>	40.89	37.68	44.41	34.81	39.42	33.26	30.28	28.68
<i>NO2</i>	33.73	28.73	26.52	23.27	25	27.32	20.79	19.76
<i>O3</i>	29.68	24.7	19.63	19.42	18.61	18.81	13.65	11.98
<i>SO2</i>	1.66	1.41	1.85	1.59	1.85	1.47	1.49	1.33

Molí del Sol

	<i>TrainMean</i>	<i>TestMean</i>	<i>X3h</i>	<i>LinearRegression</i>	<i>qr</i>	<i>IBkreg</i>	<i>M5p</i>	<i>RF</i>
<i>NO</i>	56.79	49.93	58.85	47.38	53.79	46.95	39.76	36.88
<i>NO2</i>	27.33	21.53	22.17	19.12	19.71	19.82	15.79	14.35
<i>O3</i>	33.99	22.24	19.74	19.11	18.67	16.95	13.02	11.64
<i>SO2</i>	1.97	1.52	1.6	1.4	1.6	1.62	1.45	1.38

Bulevar Sur

	<i>TrainMean</i>	<i>TestMean</i>	<i>X3h</i>	<i>LinearRegression</i>	<i>qr</i>	<i>IBkreg</i>	<i>M5p</i>	<i>RF</i>
<i>NO</i>	58.79	53.15	61.99	47.31	53.01	47.27	40.43	37.1
<i>NO2</i>	24.93	23.35	22.87	16.94	17.29	18.86	18.01	14.02
<i>O3</i>	34.34	24.05	20.34	18.48	18.16	19.43	15.21	13.75
<i>SO2</i>	1.62	1.61	1.78	1.52	1.67	1.97	1.48	1.52

Pista de Silla

	<i>TrainMean</i>	<i>TestMean</i>	<i>X3h</i>	<i>LinearRegression</i>	<i>qr</i>	<i>IBkreg</i>	<i>M5p</i>	<i>RF</i>
<i>NO</i>	35.65	34.12	40.42	32.2	35.26	30.69	31.13	26.12
<i>NO2</i>	23.62	22.13	20.96	18.42	18.98	20.56	15.89	14.14
<i>O3</i>	35.74	32.33	24.91	23.62	22.79	24.26	18.04	16.53
<i>SO2</i>	1.12	1.01	1.22	1.03	1.22	1.23	1.02	1.04

Viveros

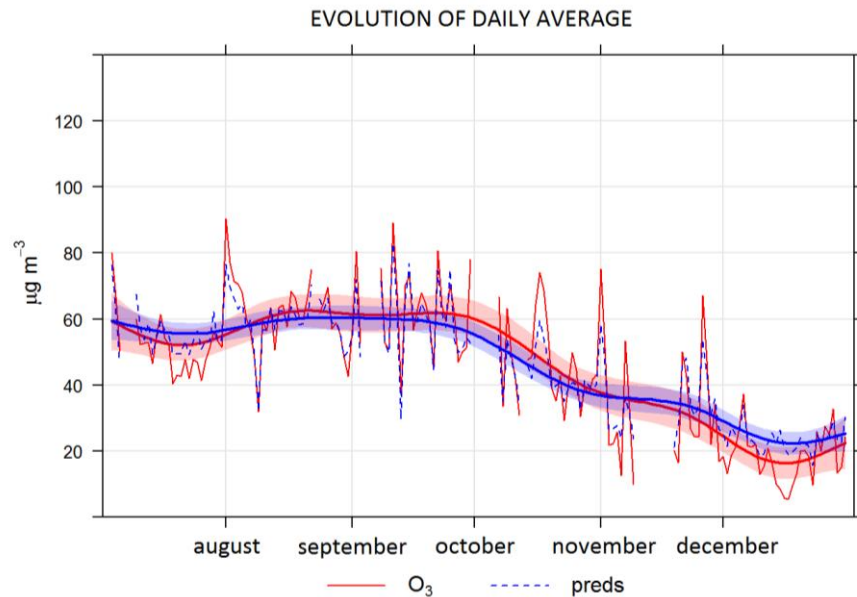
	<i>TrainMean</i>	<i>TestMean</i>	<i>X3h</i>	<i>LinearRegression</i>	<i>qr</i>	<i>IBkreg</i>	<i>M5p</i>	<i>RF</i>
<i>NO</i>	62.41	57.86	69.51	55.46	59.88	53.22	48.8	47.36
<i>NO2</i>	47.54	37.22	38.98	32.89	35.36	34.25	27.28	25.36
<i>O3</i>	37.17	25.41	21.91	22.89	22.27	20.53	15.57	13.74
<i>SO2</i>	1.62	1.37	1.72	1.39	1.71	1.57	1.61	1.31

UPV

	<i>TrainMean</i>	<i>TestMean</i>	<i>X3h</i>	<i>LinearRegression</i>	<i>qr</i>	<i>IBkreg</i>	<i>M5p</i>	<i>RF</i>
<i>NO</i>	58.09	53.01	57.68	51.43	54.24	50.82	48.81	44.92
<i>NO2</i>	39.66	29.49	28.54	26.78	27.94	28.95	21.68	21.52
<i>O3</i>	30.06	22.7	17.48	17.44	17.23	17.46	14.05	12.06
<i>SO2</i>	3.25	2.88	2.82	2.47	2.82	2.88	2.22	2.26

Avenida de Francia

Evaluación del modelo

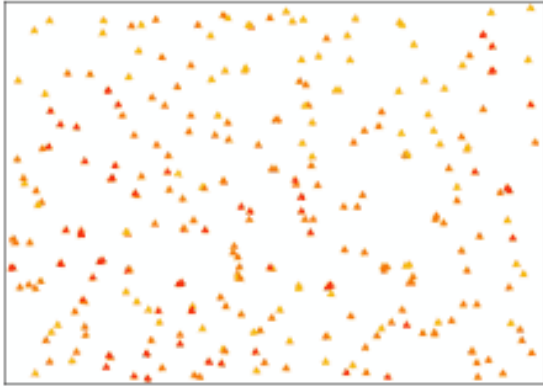


```
modStats(data, mod="preds", obs="O3")
```

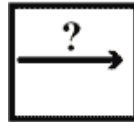
Librería: openair

Extrapolación

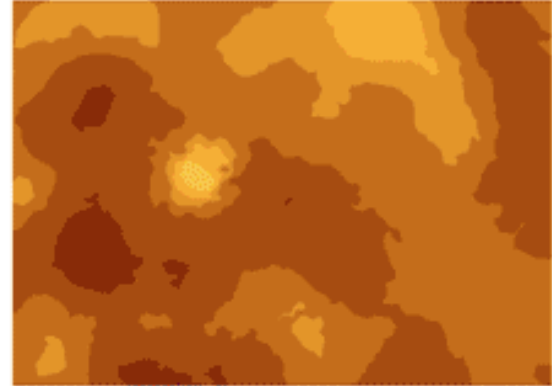
Input



Process



Output



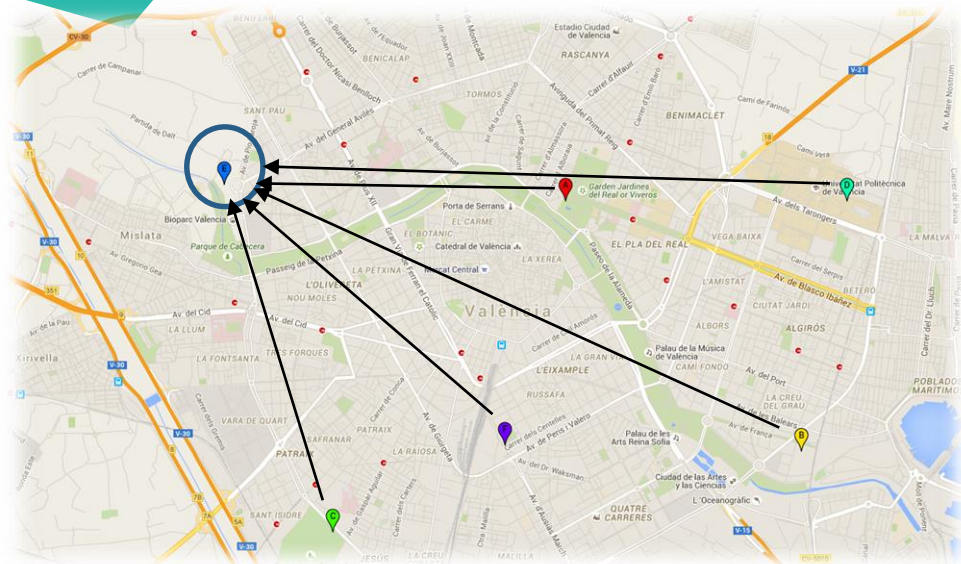
Extrapolación

Métodos de interpolación

- Media (Mean)
- Inverse Distance Weighting (IDW) con $p=1$
- Local Inverse Distance Weighting (LIDW)
- Wind Sensitive LIDW
- Ordinary Kriging

Validación cruzada

- 3 estaciones como puntos conocidos, frente a 3 estaciones como puntos desconocidos (20 posibles combinaciones).
- 4 estaciones como puntos conocidos, frente a 2 estaciones como puntos desconocidos (15 posibles combinaciones).
- 5 estaciones como puntos conocidos, frente a 1 estación como punto desconocido (5 posibles combinaciones).



Librería: geOR

```
coordenadas <- as.geodata(coordenadas)
kriging <- kslines(coordenadas,
  cov.model="exp", cov.pars=c(10,3.33), nugget=0,
  locations=c(longitud, latitud) )
prediccion<-kriging$predict
```

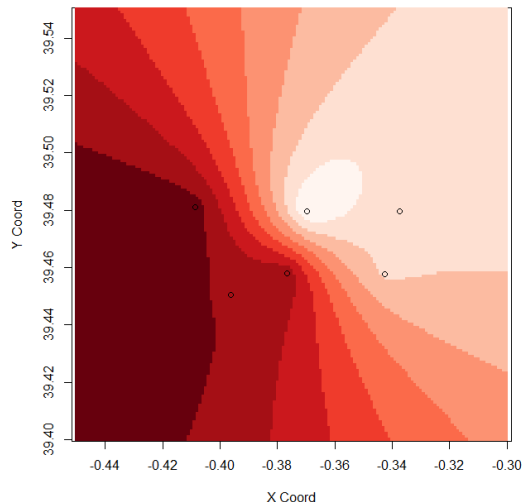
Extrapolación

Resultados: RMSE de los métodos

	3 puntos conocidos - 3 puntos desconocidos					4 puntos conocidos - 2 puntos desconocidos					5 puntos conocidos - 1 punto desconocido				
	Mean	IDW	LIDW	Wind	Krig.	Mean	IDW	LIDW	Wind	Krig.	Mean	IDW	LIDW	Wind	Krig.
NO	22.2	17.7	19.6	16.9	19.9	18.1	17.9	17.8	15.0	19.9	17.1	14.8	16.7	14.7	23.0
NO2	21.5	21.5	22.1	20.0	20.0	19.0	20.8	22.2	17.3	20.6	17.9	16.3	19.2	16.1	24.2
O3	15.4	17.5	14.5	17.7	13.2	12.3	14.4	20.4	15.1	12.7	12.0	13.9	13.7	13.9	18.7
SO2	3.1	3.0	3.2	2.9	2.9	3.2	3.0	3.0	2.5	3.4	3.0	2.4	2.9	2.5	3.4

Comparación de los cinco métodos de interpolación espacial

Extrapolación



Librería: geoR

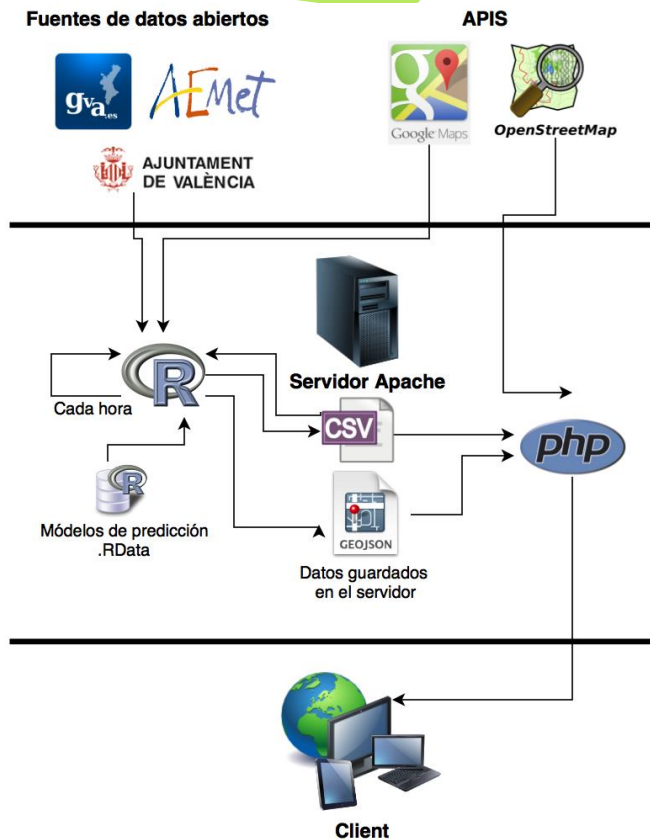
```
grd <- expand.grid(x=seq(from=x.range[1], to=x.range[2],  
  by=0.001), y=seq(from=y.range[1], to=y.range[2],  
  by=0.001))  
q <- kslide(NO, cov.model="exp", cov.pars=c(10,3.33),  
  nugget=0, locations=grd)  
image(q, val=q$predict)  
points(NO)
```



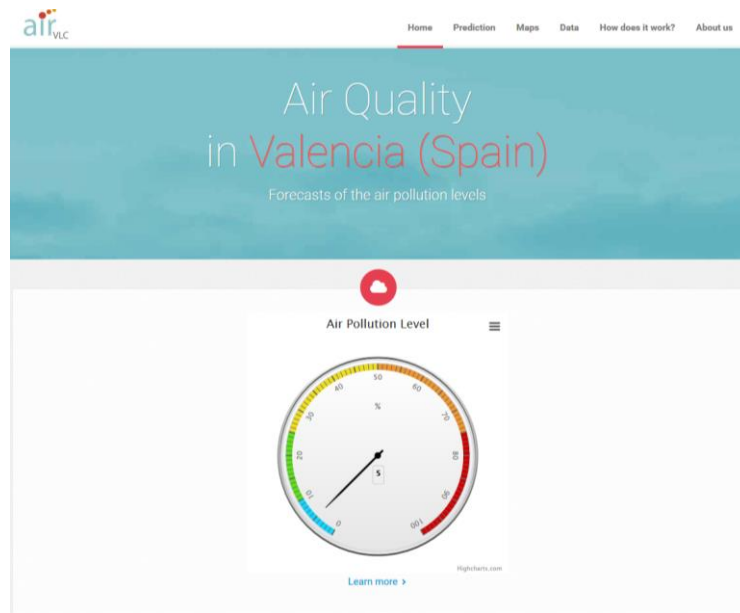
Librería: ggmap

```
Valencia <- get_map(location=c(lon=-0.3783341, lat=39.4732093),  
  zoom=14, maptype="terrain", language="es-ES")  
ggmap(Valencia, extent = "device") +  
  stat_density2d(data = coordenadas, aes(x = V2, y = V1,  
    fill = ..level.., alpha = 10), size = 10, bins = 8, geom = "polygon")  
+  
  scale_alpha(range = c(0, 0.6), guide = FALSE)+  
  scale_fill_distiller(palette = "Spectral", trans =  
    "reverse") +  
  theme_nothing(legend = FALSE))
```


Arquitectura del sistema



airVLC



<http://safe-tools.dsic.upv.es/airvlc>

Publicaciones y premios



- 3er premio H4G Valencia
- Premio Telefónica Think Big
- Premio mejor reto Smart Cities



- Premio al mejor TFM 2016 (Cátedra de transparencia y participación)
- Segundo premio Sandalio Miguel – María Aparicio 2016 a carencias y necesidades de la ciudad de Valencia (Fundación Domus)

<https://riunet.upv.es/handle/10251/71607>



Workshop Mining Urban Data: <https://goo.gl/NFIsNF>



<https://goo.gl/GjS58F>



Workshop Women in Machine Learning (póster)



Demo de la aplicación:
<http://safe-tools.dsic.upv.es/airvlc>

iGracias! =)

liconoc@upv.es
@liconoc