Modelling and Forecasting Air Quality data with missing values via multivariate time series: Application to Madrid

POLITÉCNICA

INDUSTRIALES
ETSII | UPM

Authors: Mario Ramírez Jaén, Carolina García Martos & Mª Jesús Sánchez Naranjo

July 2017

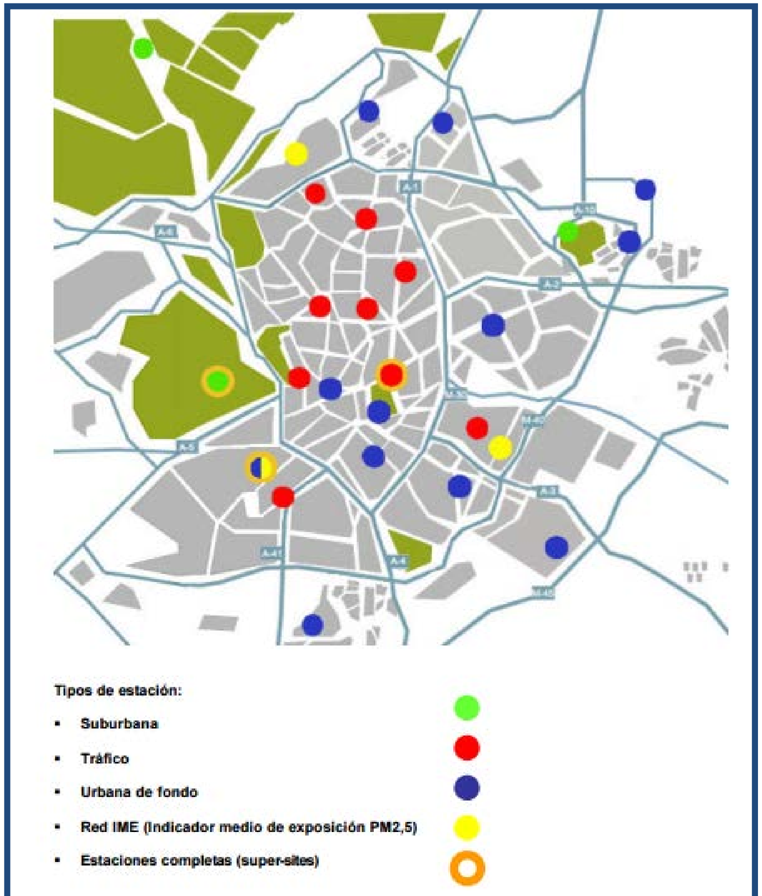
Context

Big cities suffer from Air Quality Issues.

New regulations show up at both national and European scope.

The use of monitorization networks to measure pollution and model as well as forecasting is spreading (Lawson et al. (2011), Febrero-Bande et al. (2007), Castellano et al. (2009)).

Madrid has a network made of **24 stations**. They are the following:



Pza. del Carmen	**Barajas**
Pza. de España	Méndez Álvaro
Barrio del Pilar	Castellana
Escuelas Aguirre	**Retiro Park**
Cuatro Caminos	Pza. Castilla
Av. Ramon y Cajal	Ensanche Vallecas
Vallecas	Urb. Embajada
Arturo Soria	Pza. Fdez. Ladrea
Villaverde Alto	Sanchinarro
C/ Farolillo	**El Pardo**
Moratalaz	Parque Juan Carlos I
Casa de Campo	Tres Olivos

Steps followed along this work-in-progress Project:

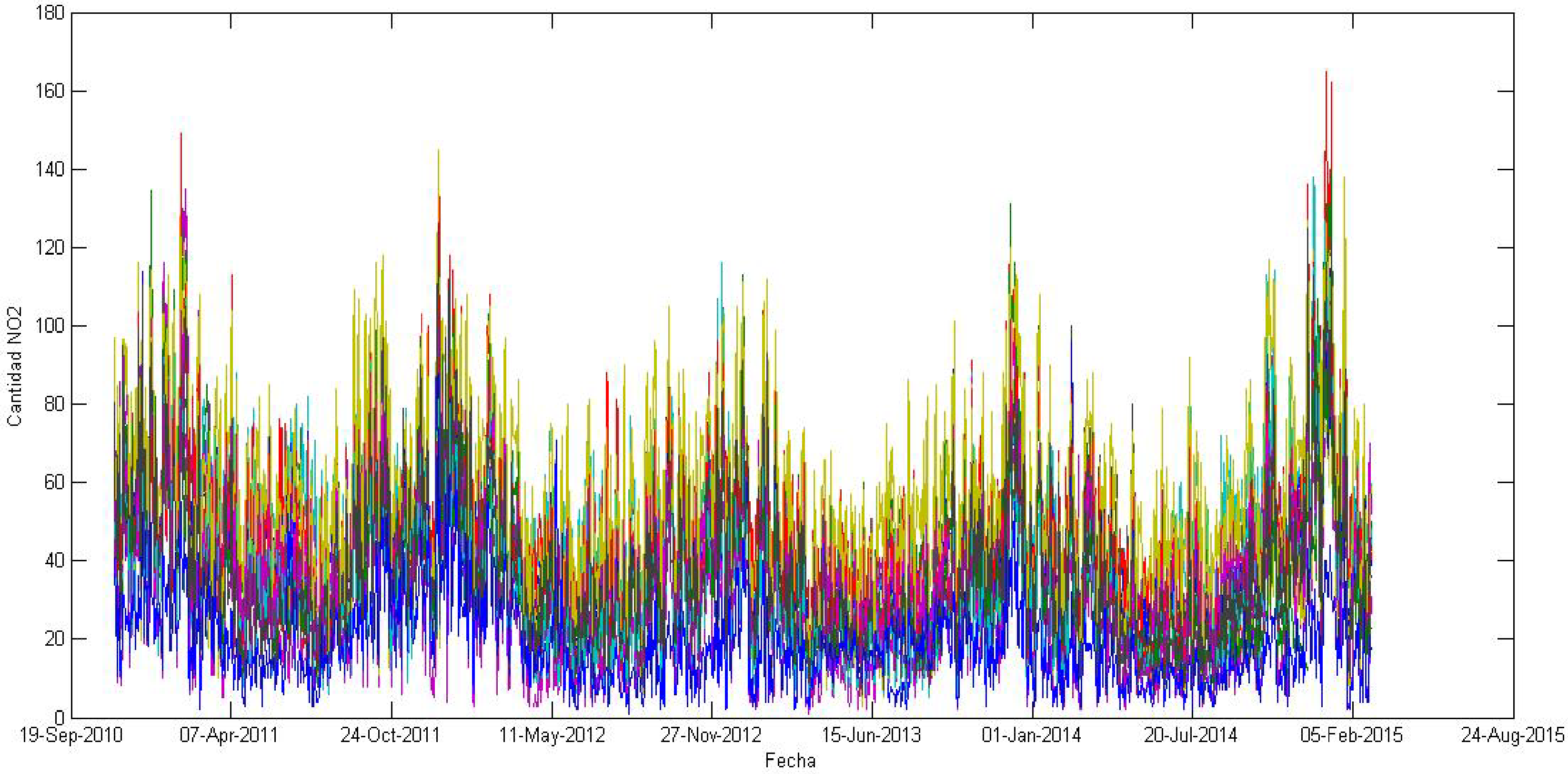
1. Daily data:
   - Univariate modelling of daily data, for each of the series corresponding to 22 stations in the city of Madrid.
   - DFM for the 22-dimensional vector of series.

2. Hourly data:
   - DFM for the 24 hourly series corresponding to a single station: weekly and yearly seasonality.
   - DFM for the 24·m series, where m corresponds to the seleted number of stations to be analyzed.

All the aforementioned approaches need intervention of the missing data, and then working with the “corrected series”.

State-space formulation with missing data.

Brief descriptive analysis of the daily data: NO_2, 22 stations

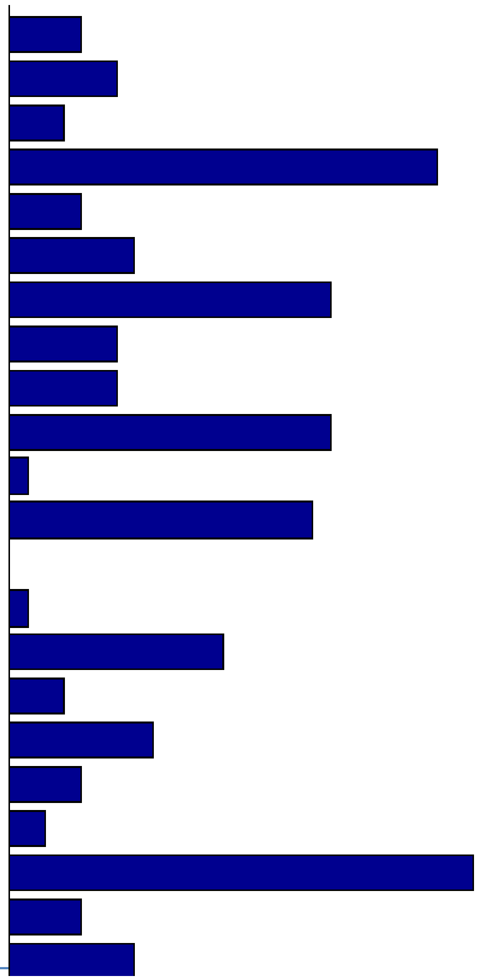


✓ **Relationship between these series**

✓ **Several seasonalities**

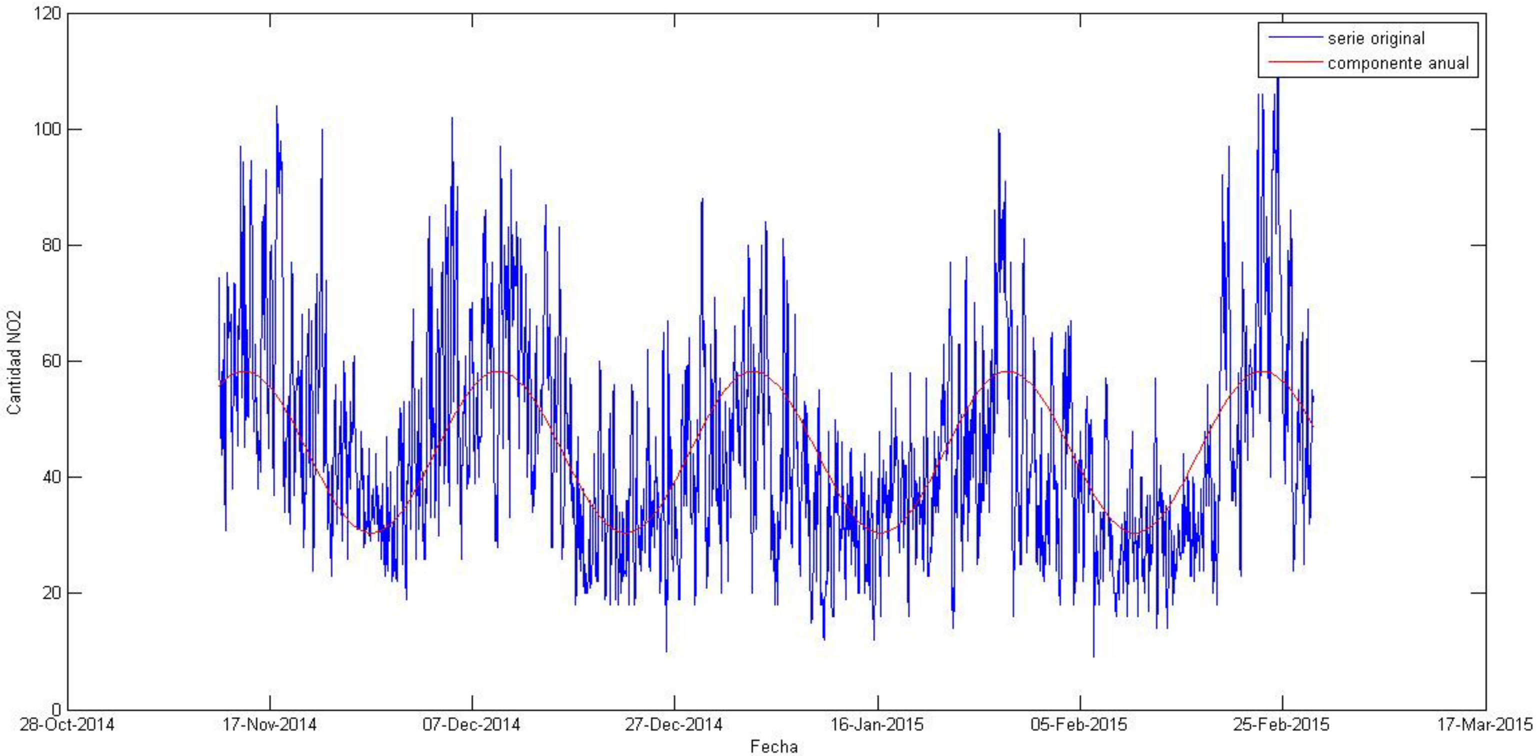
Number of missing data per station

STATION	Missing data
Pza. del Carmen	4
Pza. de España	6
Barrio del Pilar	3
Escuelas Aguirre	24
Cuatro Caminos	4
Av. Ramón y Cajal	7
Vallecas	18
Arturo Soria	6
Villaverde Alto	6
C/ Farolillo	18
Moratalaz	1
Casa de Campo	17
Barajas	0
Méndez Álvaro	1
Castellana	12
Retiro Park	3
Pza. Castilla	8
Ensanche Vallecas	4
Urb. Embajada	2
Pza. Fdez. Ladrea	26
Sanchinarro	4
El Pardo	7



Year seasonality treatment

Used Software



$$Year\ deterministic\ seasonality = a_1 + b_1 sin\left(\frac{2\pi t}{365}\right) + c_1 cos\left(\frac{2\pi t}{365}\right)$$

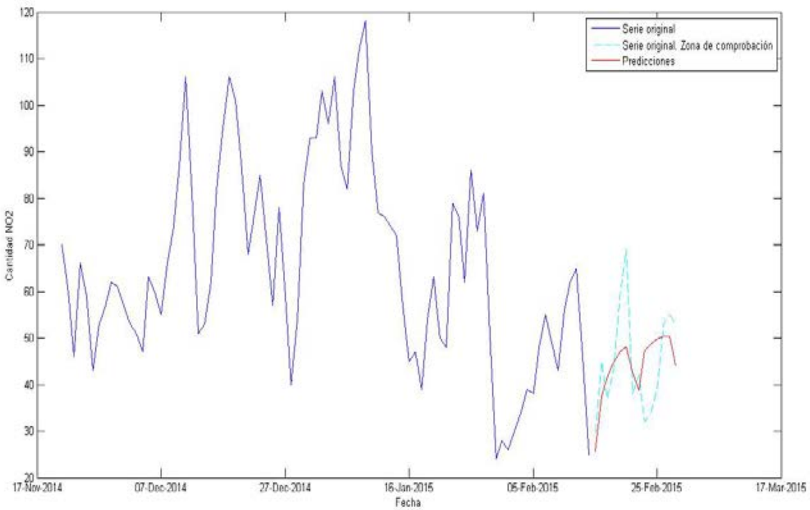
First approach: SARIMA :$(p,d,q)\times(bp,bd,bq)_s$ models for daily NO_2 concentrations

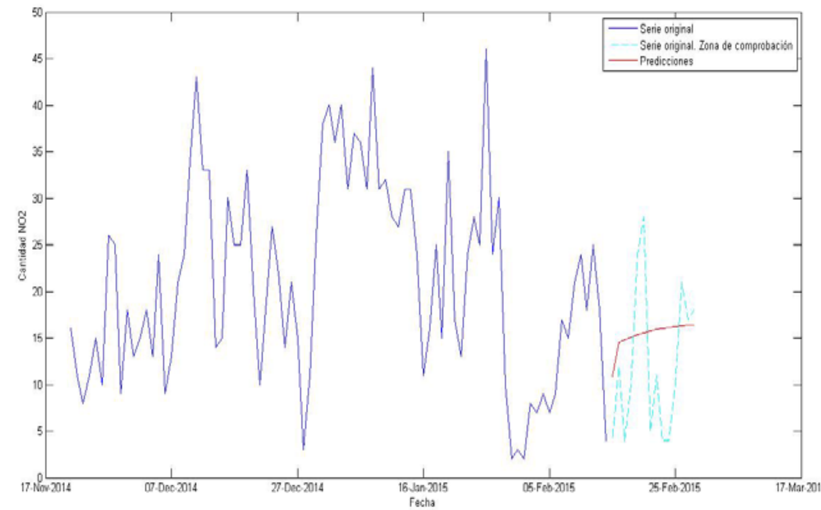
	Pza. del Carmen	Pza. de España	Barrio del Pilar	Escuelas Aguirre	Cuatro Caminos	Ramón y Cajal	Vallecas	Arturo Soria	Villaverde Alto	C/ Farolillo	Moratalaz
p	1	1	1	1	1	1	1	1	1	1	1
d	0	0	0	0	0	0	0	0	0	0	0
q	0	0	0	0	0	0	0	0	0	0	0
bp	0	0	0	0	0	0	0	0	0	0	0
bd	1	1	1	1	1	1	1	1	1	1	1
bq	1	1	1	1	1	1	1	1	1	1	1

	Casa de Campo	**Barajas**	Mendez Alvaro	Castellana	**Retiro**	Pza. Castilla	Ensanche Vallecas	Urb. Embajada	Pza. Fdez. Ladrea	Sanchinarro	**El Pardo**
p	**1**	**1**	1	1	**1**	1	1	1	1	1	**1**
d	**0**	**1**	0	0	**0**	0	0	0	0	0	**1**
q	**2**	**1**	0	0	**2**	0	0	0	0	0	**1**
bp	**0**	**0**	0	0	**0**	0	0	0	0	0	**0**
bd	**0**	**1**	1	1	**1**	1	1	1	1	1	**1**
bq	**0**	**1**	1	1	**1**	1	1	1	1	1	**1**

Forecasting Error

$$Error = \frac{|\hat{y}_t - y_t|}{y_t}$$



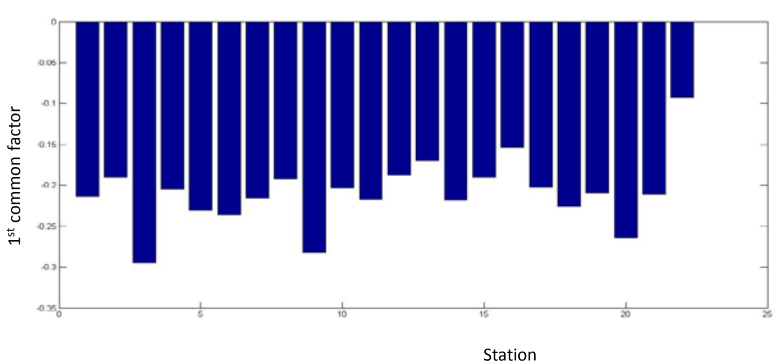


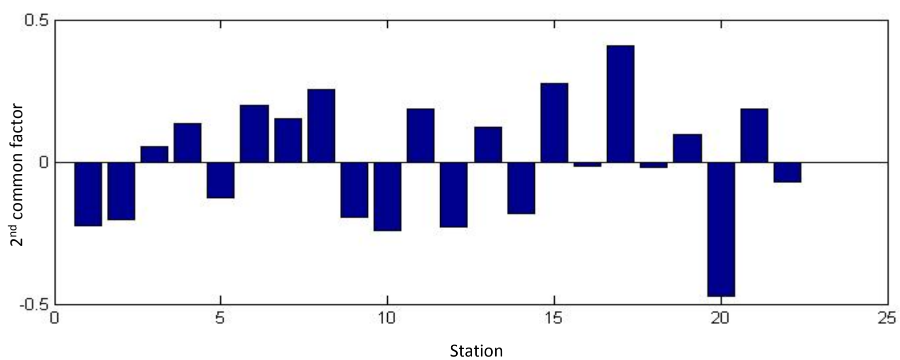
1	2	3	4	5	6	7	8	9	10	11
12%	4%	19%	3%	13%	6%	7%	17%	7%	34%	12%
12	***13***	14	15	16	17	18	19	20	21	***22***
215%	***75%***	5%	9%	5%	17%	14%	44%	23%	43%	***171%***

Used Software

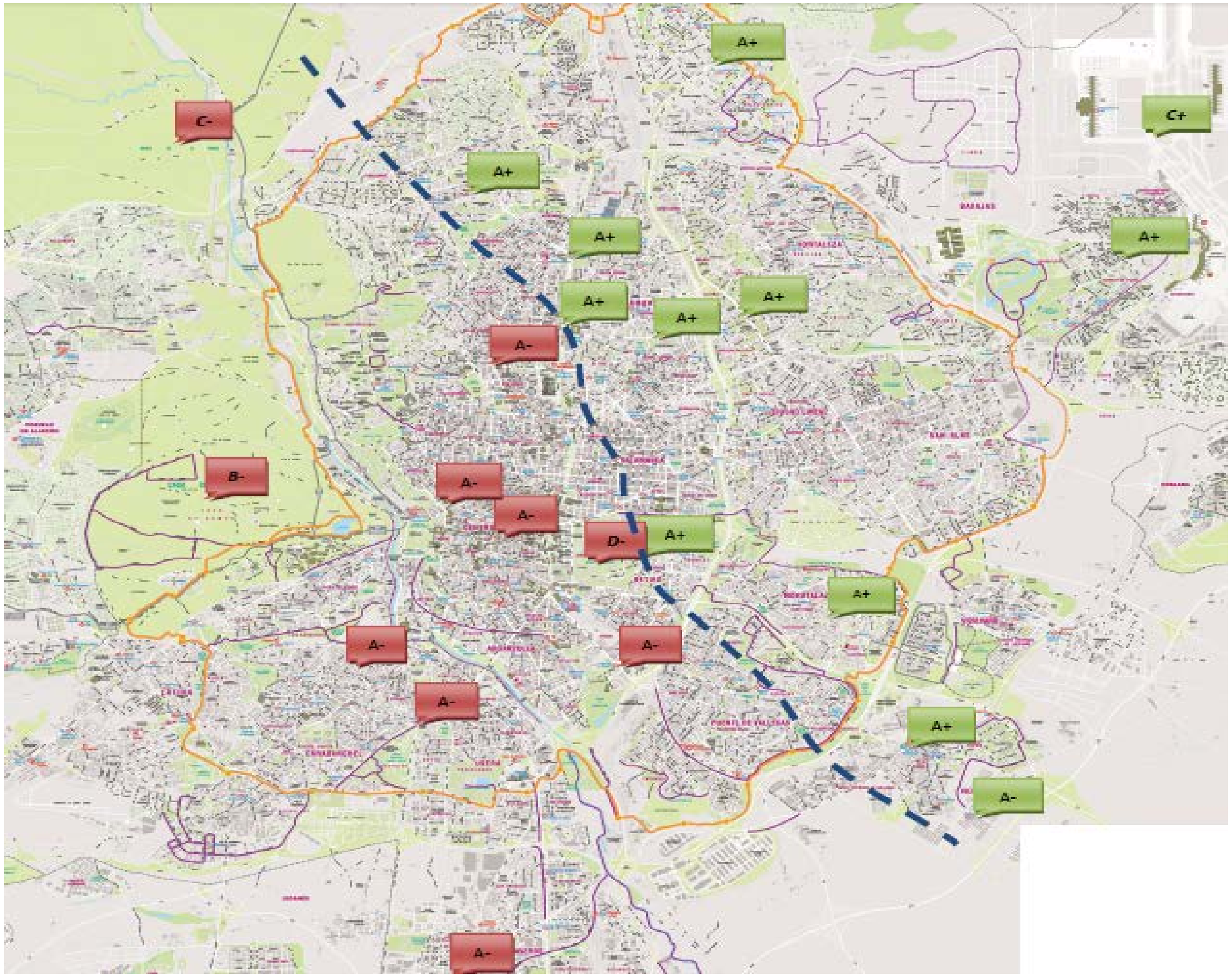
“An amount of unobserved common factors (r), clearly smaller than the amount of series (r<<m), can explain the series variability”.

1	2	3	4	5	6	7	8	9	10	11
84,7%	4,1%	1,7%	1,6%	1,0%	0,9%	0,8%	0,7%	0,6%	0,5%	0,4%
12	13	14	15	16	17	18	19	20	21	22
0,4%	0,4%	0,3%	0,3%	0,3%	0,3%	0,2%	0,2%	0,2%	0,1%	0,1%



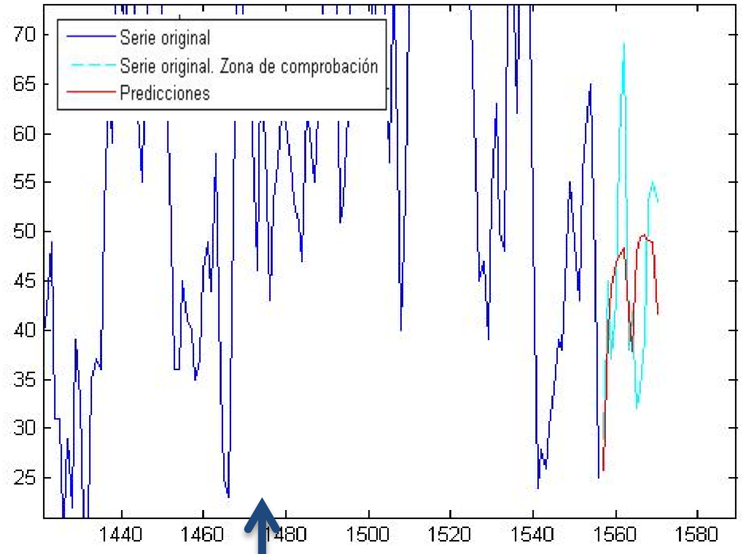


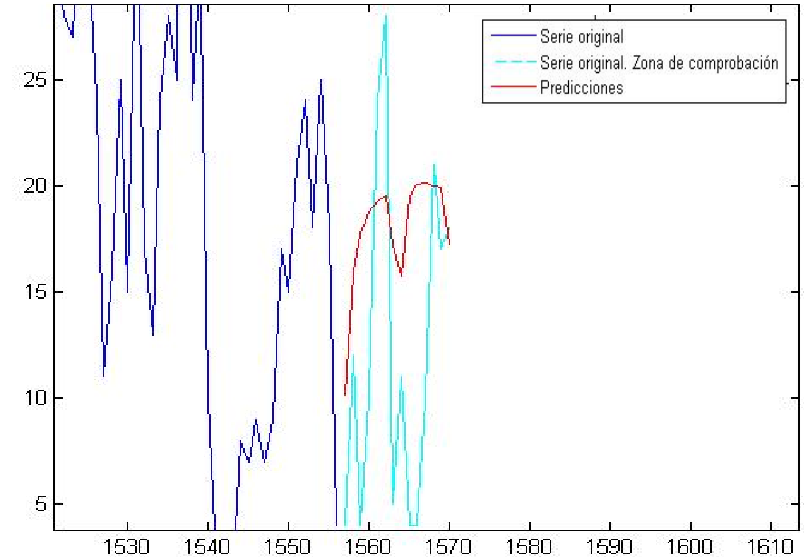
Geographical interpretation of the 2nd common factor



Forecasting Error

$$Error = \frac{|\hat{y}_t - y_t|}{y_t}$$





1	2	3	4	5	6	7	8	9	10	11
18%	21%	36%	21%	40%	18%	27%	26%	31%	46%	23%
12	13	14	15	16	17	18	19	20	21	***22***
92%	26%	37%	27%	26%	21%	37%	26%	38%	21%	***75%***

¿Why?

Regulations establishes hourly pollution limits.

Challenges

- Course of dimensionalty
- Double Seasonality

Solutions

- Select just two stations
- SCA (software)

Approaches

- Parallel approach station by station
- Parallel approach for all series

Parallel Approach Station by Station

Used Software

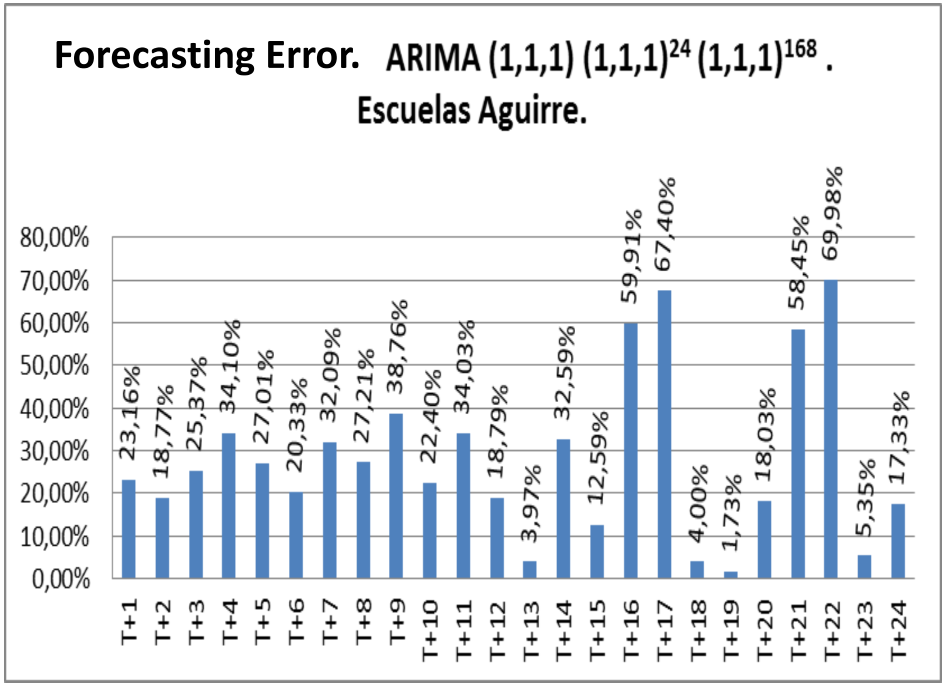
SCA

¡Double Seasonality!

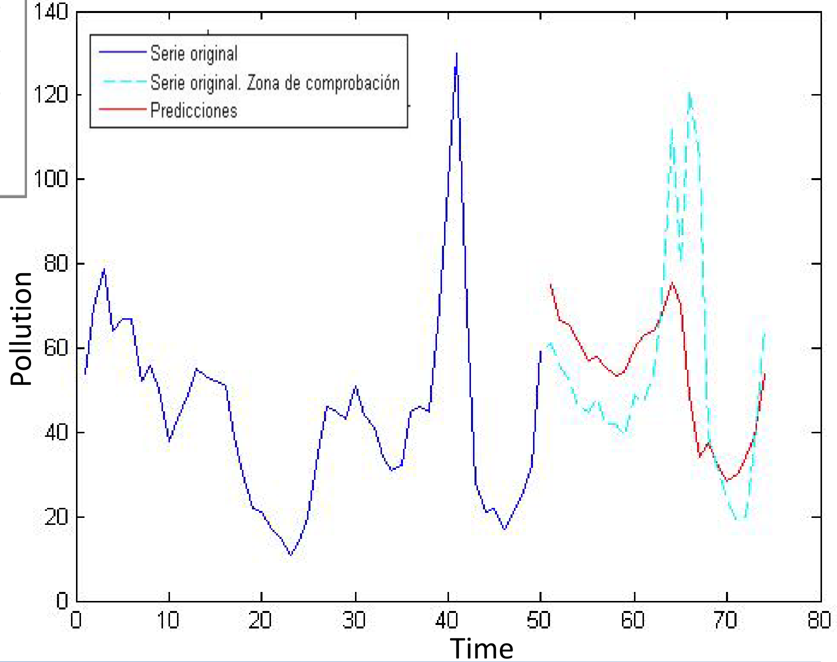
$$\text{ARIMA }(1,1,1)\text{x}(1,1,1)_{24}\text{x}(1,1,1)_{168}$$

```
---------------------------------------------------------------------
VARIABLE    TYPE OF     ORIGINAL      DIFFERENCING
            VARIABLE   OR CENTERED
                                          1       24      168
ESCUELAS    RANDOM      ORIGINAL      (1-B  ) (1-B  ) (1-B   )
---------------------------------------------------------------------
```

PARAMETER LABEL		VARIABLE NAME	NUM./ DENOM.	FACTOR	ORDER	CONS- TRAINT	VALUE	STD ERROR	T VALUE
1	CONST		CNST	1	0	NONE	-.0015	.0106	-.15
2	THETA1	ESCUELAS	MA	1	1	NONE	-.2859	.1180	-2.42
3	THETA24	ESCUELAS	MA	2	24	NONE	.8088	.0101	79.78
4	THETA168	ESCUELAS	MA	3	168	NONE	.6991	.0123	56.80
5	PHI1	ESCUELAS	AR	1	1	NONE	-.2134	.1207	-1.77
6	PHI24	ESCUELAS	AR	2	24	NONE	-.0630	.0169	-3.73
7	PHI168	ESCUELAS	AR	3	168	NONE	-.2895	.0161	-18.03



Forecasts become inaccurate after 12 hours, h>12.



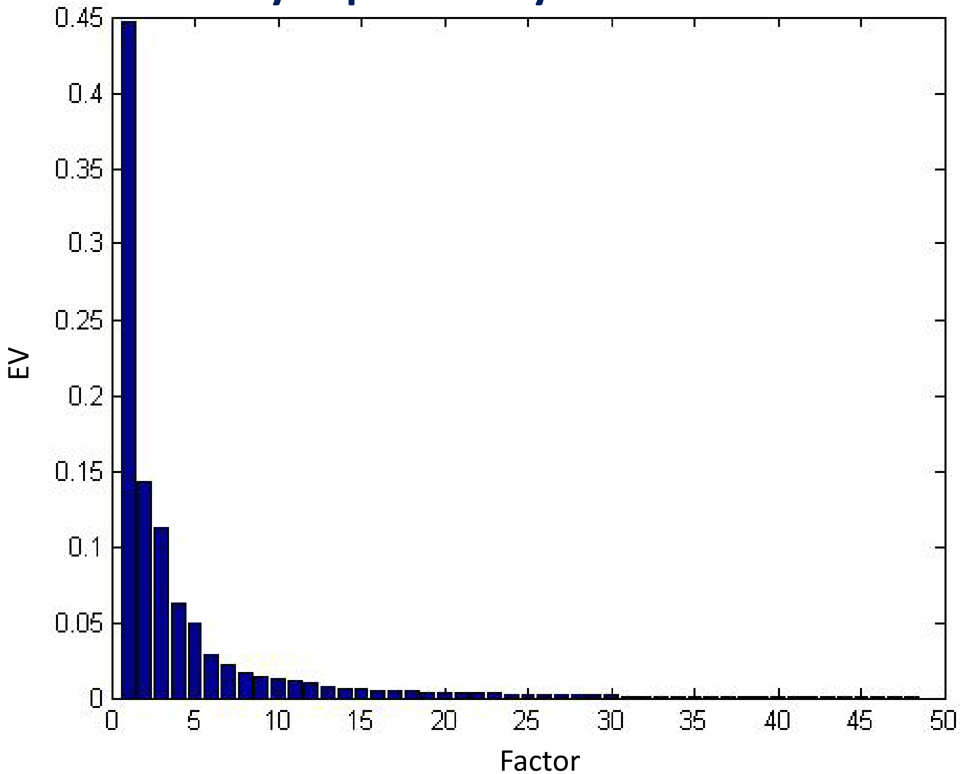
Forecasting error around 20% for the first 12 hours (forecasting horizon h=1,…,12).

Parallel Approach All Stations

Used Software

Dynamic Factor Model (DFM)

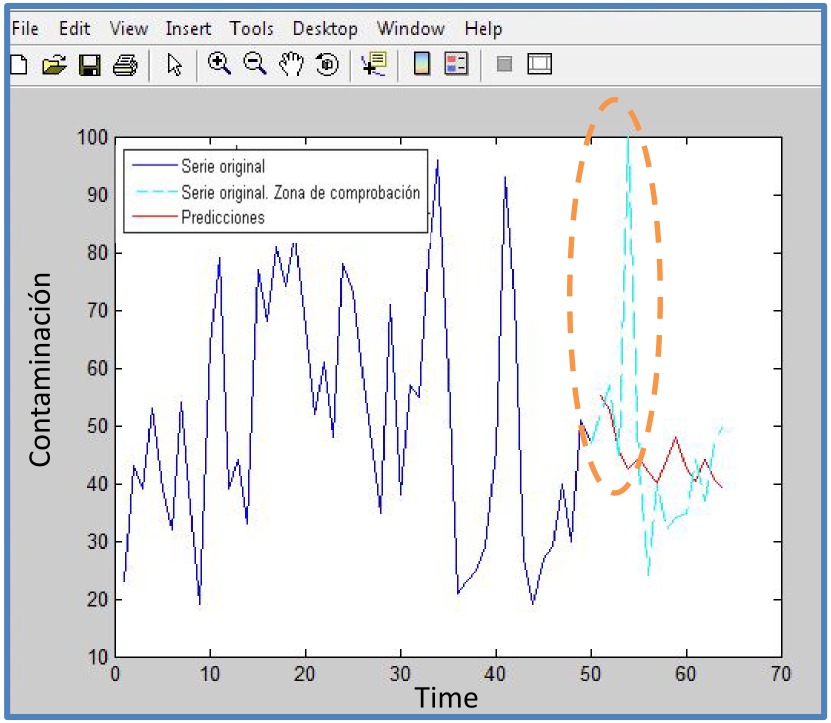
Variability explained by the common factors



3 Factors

	Factor 1	Factor 2	Factor 3
p	1	0	1
d	0	0	0
q	0	1	0
bp	0	0	0
bd	1	0	1
bq	1	0	1

Average forecasting error around 40%. Great amount of outliers

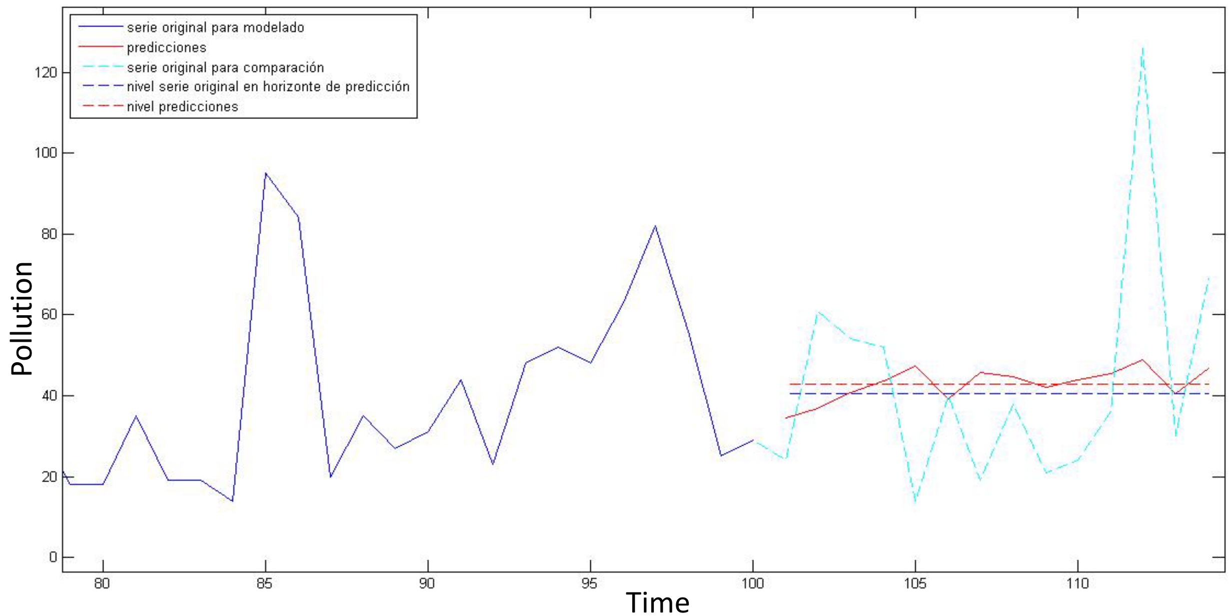


Example: Window 505, hour 1, Escuelas Aguirre

Horizon	1	2	3	4	5
Error	5,92%	7,3%	1,19%	57,42%	3%

(Sep-Oct 2013)

Forecasting Error when trying to forecast the level of the series



Hora 1	Hora 2	Hora 3	Hora 4	Hora 5	Hora 6	Hora 7	Hora 8	Hora 9	Hora 10	Hora 11	Hora 12
25,48%	20,14%	18,23%	19,66%	18,24%	15,36%	14,72%	15,10%	14,44%	15,20%	16,34%	18,06%
Hora 13	Hora 14	Hora 15	Hora 16	Hora 17	Hora 18	Hora 19	Hora 20	Hora 21	Hora 22	Hora 23	Hora 24
15,29%	13,65%	15,51%	15,42%	13,98%	13,39%	11,66%	12,98%	10,54%	8,86%	11,00%	13,71%

INDUSTRIALES
ETSII | UPM

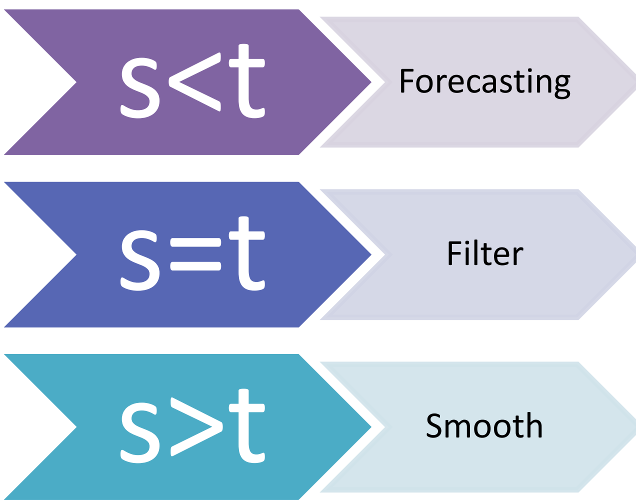
State Equation $\boldsymbol{x}_t = \phi \boldsymbol{x}_{t-1} + \boldsymbol{w}_t$

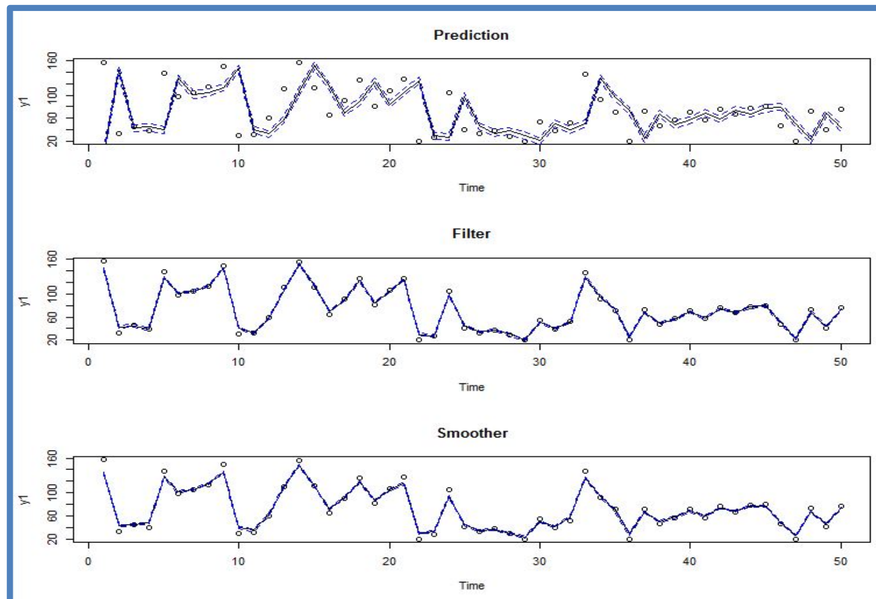
Observation Equation $\boldsymbol{y}_t = A_t \boldsymbol{x}_t + \boldsymbol{v}_t$

Used Software

R

Aim: Estimate the value of x_t from the observations $Y_s = \{y_1, \dots, y_s\}$





SS with Missing Data

Missing data do not modify the problems dimension

$$\begin{pmatrix} y_t^{(1)} \\ y_t^{(2)} \end{pmatrix} = \begin{bmatrix} A_t^{(1)} \\ A_t^{(2)} \end{bmatrix} x_t + \begin{pmatrix} v_t^{(1)} \\ v_t^{(2)} \end{pmatrix}$$

Being $y_t^{(2)}$ missing data

$$y_t = \begin{pmatrix} y_t^{(1)} \\ 0 \end{pmatrix}; \qquad A_t = \begin{bmatrix} A_t^{(1)} \\ 0 \end{bmatrix}; \qquad R_t = \begin{bmatrix} R_{11t} & 0 \\ 0 & I_{22t} \end{bmatrix}$$

Kalman Filter

$$x_t^{t-1} = \phi x_{t-1}^{t-1} + w_t$$
$$P_t^{t-1} = \phi P_{t-1}^{t-1} \phi' + Q$$
$$x_t^t = x_t^{t-1} + K_t(y_t - A_t x_t^{t-1})$$
$$P_t^t = [I - K_t A_t P_{t-1}^{t-1}] P_t^{t-1}$$

Kalman Gain: $K_t = P_t^{t-1} A_t' \left[A_t P_t^{t-1} A_t' + R\right]^{-1}$

forecasting Error: $\varepsilon_t = y_t - E(y_t | Y_{t-1}) = y_t - A_t x_t^{t-1}$

Variance-Covariance matrix: $\Sigma_t = A_t P_t^{t-1} A_t' + R$

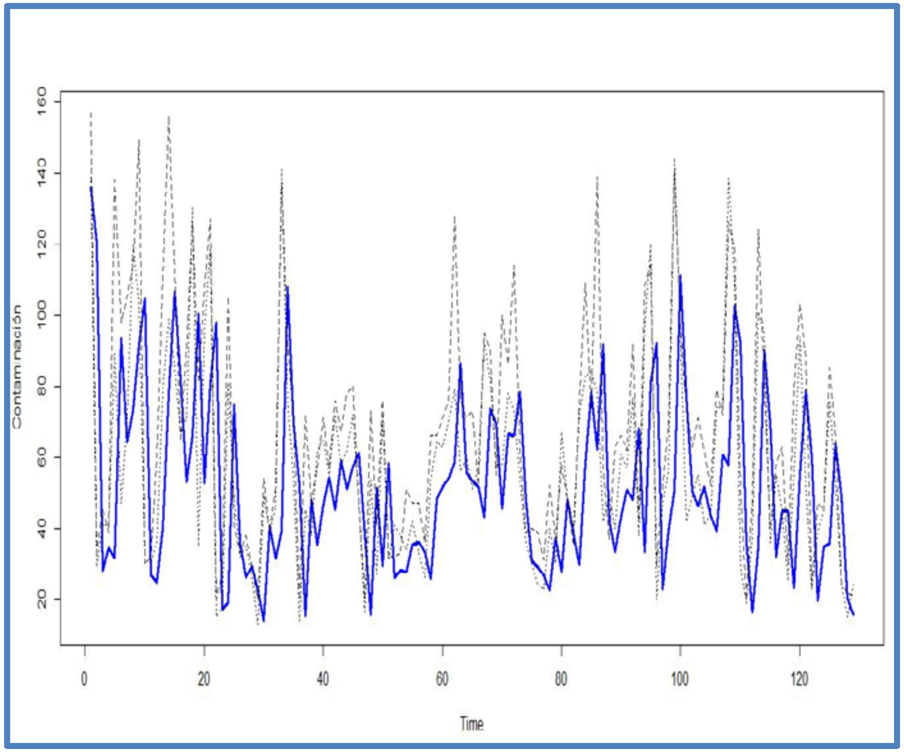
Example with two series

Escuelas Aguirre, hour 1 & hour 2

Estimated parameters:

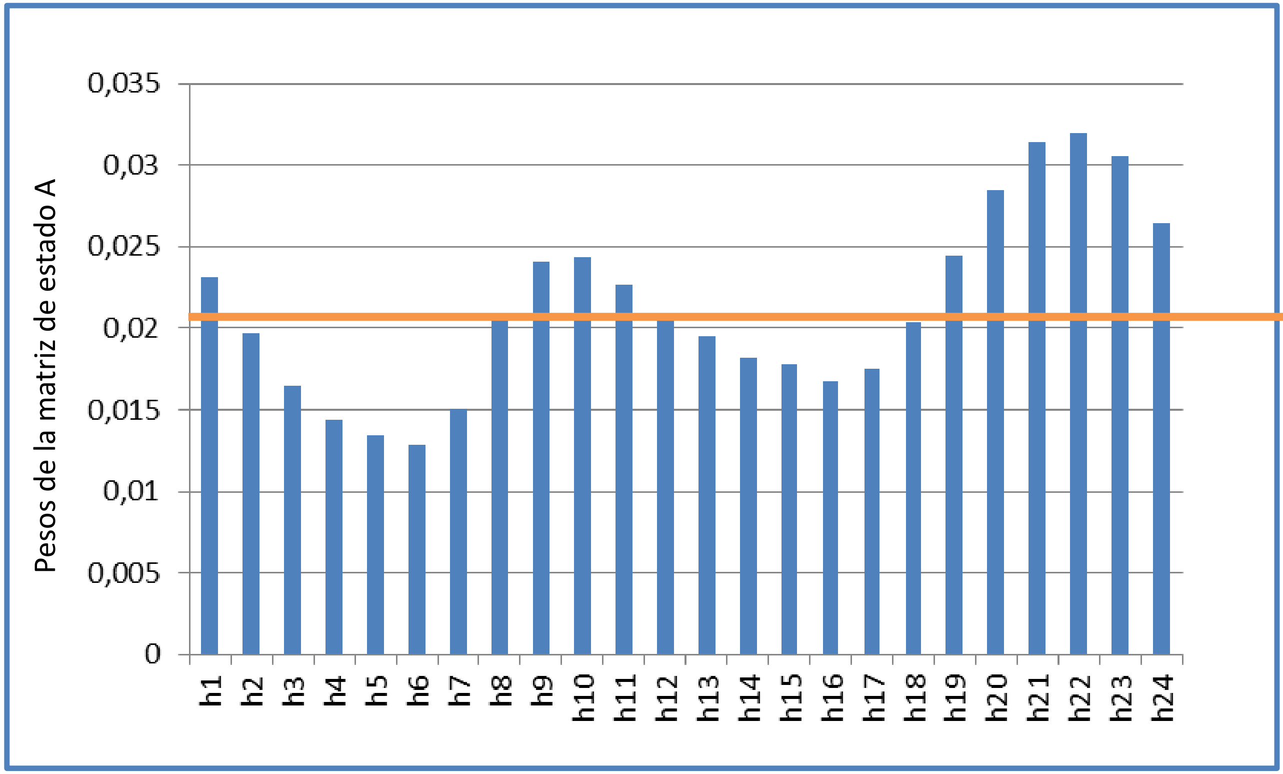
	estimate	SE
Q	30.7550812	7.5092844
R11	8.7013815	2.1849004
R22	[illegible]	1.4048085
Phi	0.8655842	0.0417215
A1	1.1483367	0.2712761
A2	0.9705190	0.2295206

~1 Similar to Bivariate Local Level Model



Example with 24 series

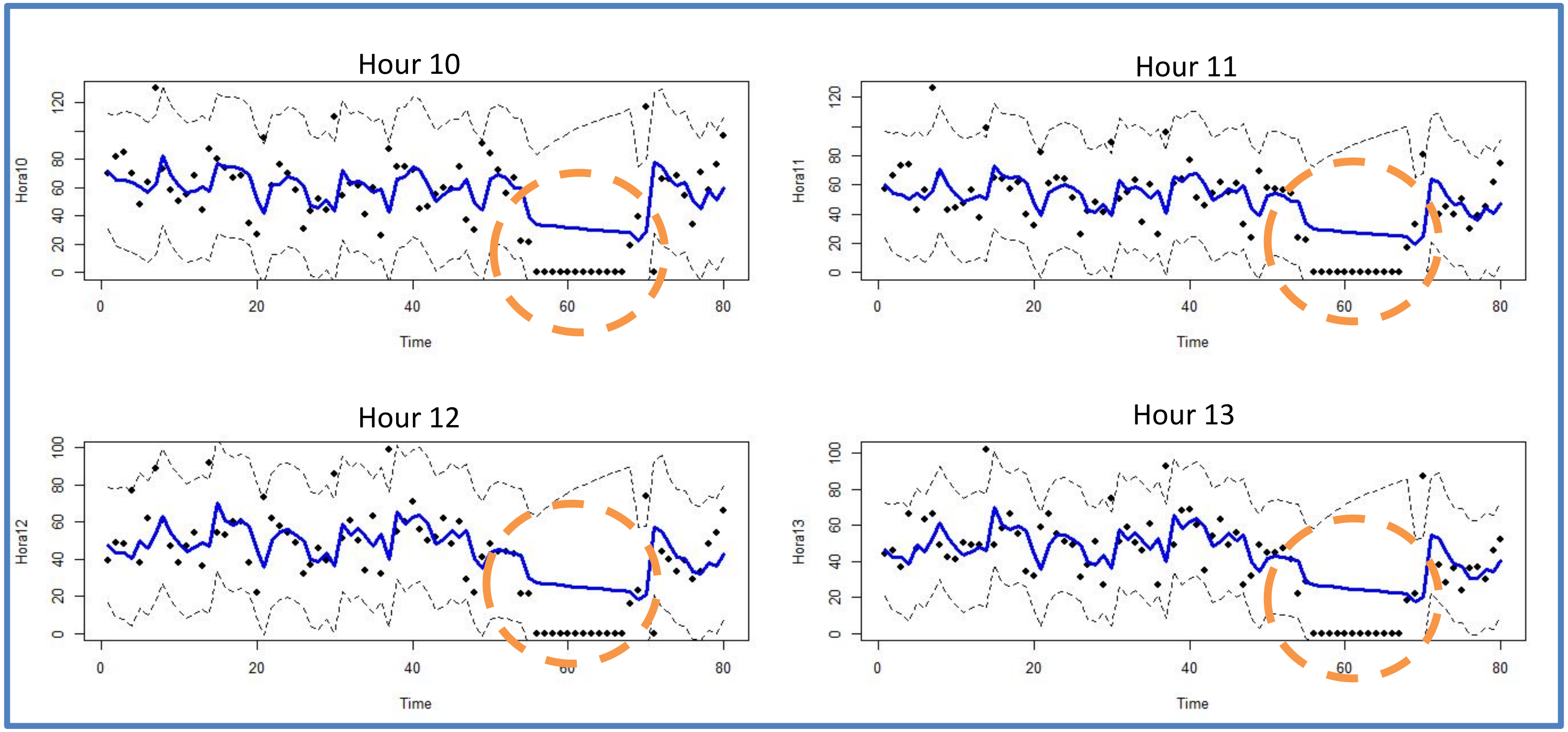
$\Phi \sim 1$ ⇨ Similar to a Multivariate Local Level



$A_i \sim 0{,}21 = 1/48$

SS with Missing Data

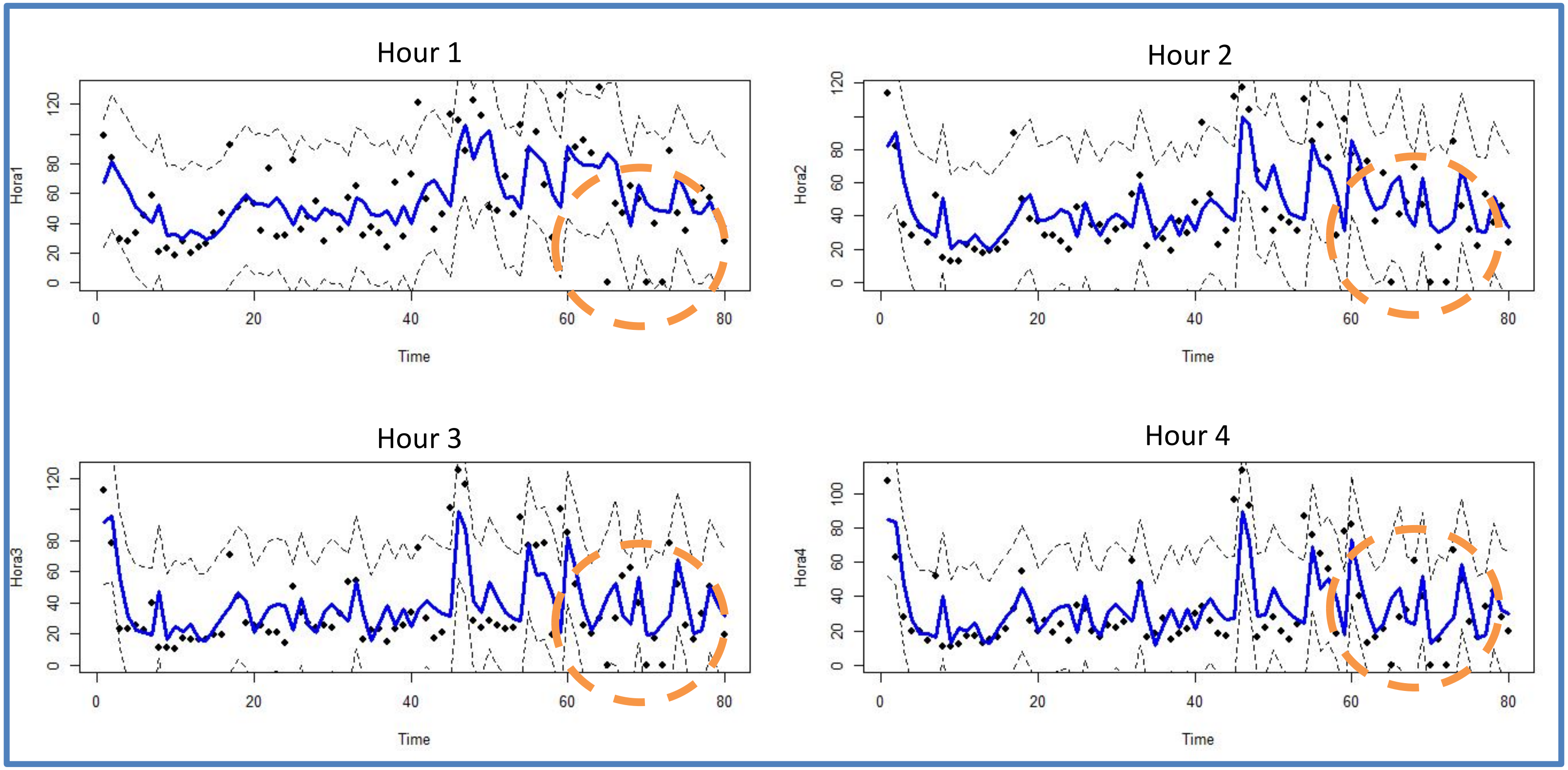
Example with consecutive missing data:



Flat shape, far from reality

SS with Missing Data

Example with isolated missing data:



Good fit to the real series

Daily data conclusions:

There is a dominant model for daily data series: ARIMA (1,0,0)×(0,1,1)

A geographic dependence, with two clearly differentiated areas, has been identified

El Pardo & La Casa de Campo can't be forecasted with an hourly data model

Hourly data conclusions:

With an ARIMA (1,1,1) × (1,1,1)24 × (1,1,1)168 20% forecasting error can be achieved for no disaggregated data.

With disaggregated date model can only make level forecasting

Hourly data are more unstable and have more outliners tan daily date, making them more difficult to forecast.

State space conclusions:

The best value for the transition matrix in close to 1, being that the value used in a Local Level model.

With few series, observation matrix components are close of those of a Local Level.

For a big number of series n, observation matrix components have a value around 1/n.

Further Research

3D Spatio-temporal modelling

Consider more general models than just a MLL

Study relationship between pollutants
and Extend the State Space formulation for the data

Forecasting with SS models

THANK YOU FOR YOUR ATTENTION

Modelling and Forecasting Air Quality data with missing values via multivariate time series: Application to Madrid

27th Annual Conference of The International Environmetrics Society
joint with GRASPA 2017 on Climate and Environment
24-26 July 2017 - Bergamo, Italy

INDUSTRIALES
ETSII | UPM

Authors: Mario Ramírez Jaén, Carolina García Martos & Mª Jesús Sánchez Naranjo

July 2017