Indian Institute of Technology, Delhi

TIME SERIES PAPER PRESENTATION

Forecast $PM_{2.5}$ Pollution in Chile

Adapted from : A dynamic multiple equation approach for forecasting PM2.5 pollution in Santiago, Chile by Stella Moisan, Rodrigo Herrera, Adam Clements

Supervised by : Prof. Sorabh Paul

Presented by : Riya Singh

Introduction

To propose method of time series forecasting which incorporates hourly, daily and seasonal characteristics for predicting hourly PM2.5 pollution concentrations

- The paper aims to forecast PM 2.5 concentrations in Santiago, Chile using dynamic multiple linear equations
- Compare accuracy with competing non linear forecasting models in terms of both fit and predictive ability
- Forecast useful for government policies, env. laws etc.

Introduction

- Chile ranked as country with highest pollution amongst 36 countries of OECD in 2014
- Santiago, capital city, has 4th worst air quality in the continent
- According to National System of Environmental Information, net economic benefit of regulating PM2.5 is estimated to be USD 7.1 billion
- Predictive models are important in order to allow government authorities to take efficient action to minimise the economic consequences of heightened pollution levels

Dataset

- Hourly Historical observations of weather & environmental concentrations
    - Collected from 11 monitoring stations in Santiago
    - Collected from National Air Quality Information System
    - 1 January 2011 - 31 August 2015
- Hourly; Weekly; Monthly
- Critical episodes
    - Alert (80-109 microgram/m^3)
    - Pre-emergency (110-169 microgram/m^3)
    - Emergency (>170 microgram/m^3)

Dataset

- Hourly average PM2.5 concentrations for the 2011–2015 period at each station
  - station with high pop. density have high concentrations
- Non uniform proximity of 11 stations :
  - Cerro Navia, Pudahuel likely to have more similarities in data
  - Talagante & Las Condes largely independent

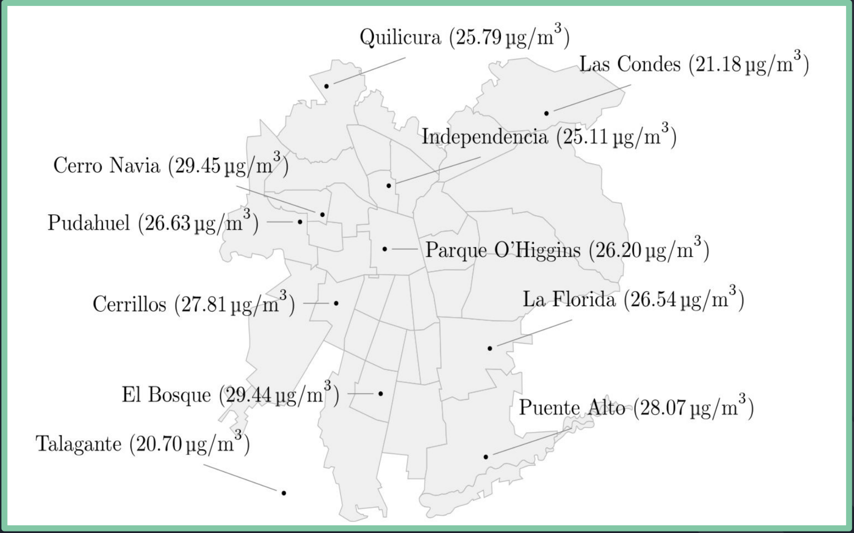

Quilicura (25.79 µg/m^3)
Las Condes (21.18 µg/m^3)
Independencia (25.11 µg/m^3)
Cerro Navia (29.45 µg/m^3)
Pudahuel (26.63 µg/m^3)
Parque O'Higgins (26.20 µg/m^3)
Cerrillos (27.81 µg/m^3)
La Florida (26.54 µg/m^3)
El Bosque (29.44 µg/m^3)
Puente Alto (28.07 µg/m^3)
Talagante (20.70 µg/m^3)

Figure 1 : Monitoring stations of Chile

Dataset

- PM2.5 concentrations at Pudahuel monitoring station
    - largest geographical area, with one of the highest pollution

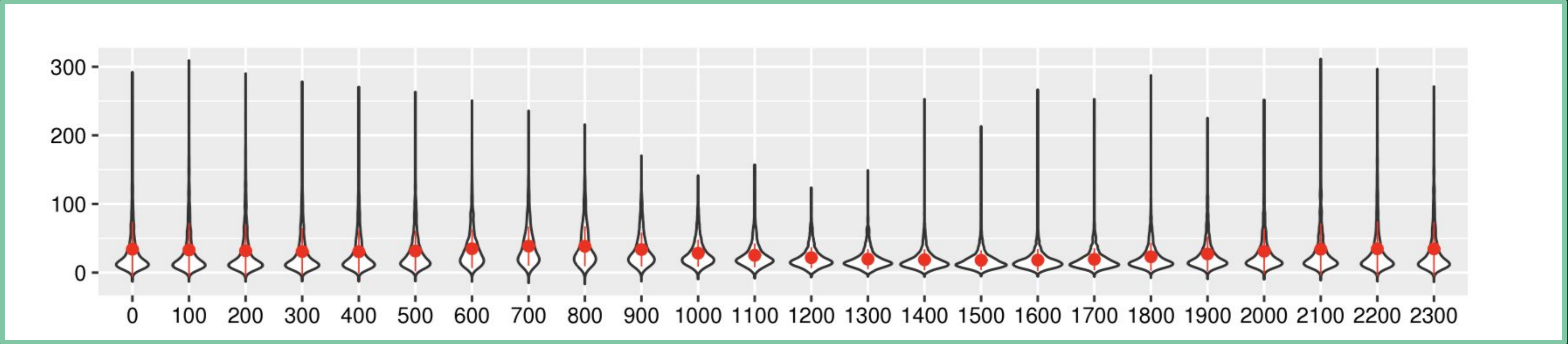


Figure 2 : Hourly average Pudahuel station January 2011 - August 2015

- 6 AM - 10 AM : highest average ; 6 PM - 9 PM : next peak in average
- Why? Work.

Dataset

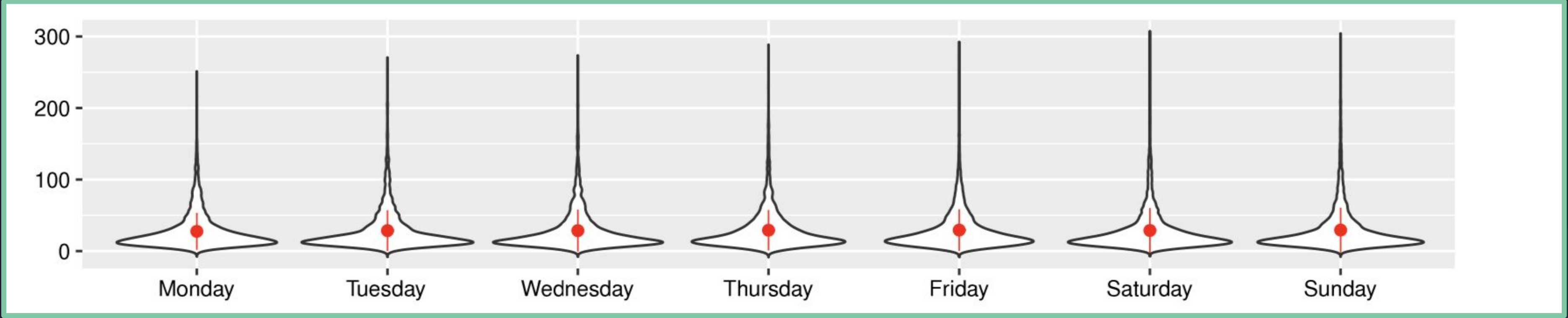


Figure 3: Weekly average, Pudahuel station January 2011 - August 2015

- Difficult to see, but weekend concentration slightly lower
- Greater dispersion observed on Friday, Saturday, Sunday
- Weekend irregularity? People returning to work [CONAMA, 2005]
- Stable pattern on weekdays

Dataset

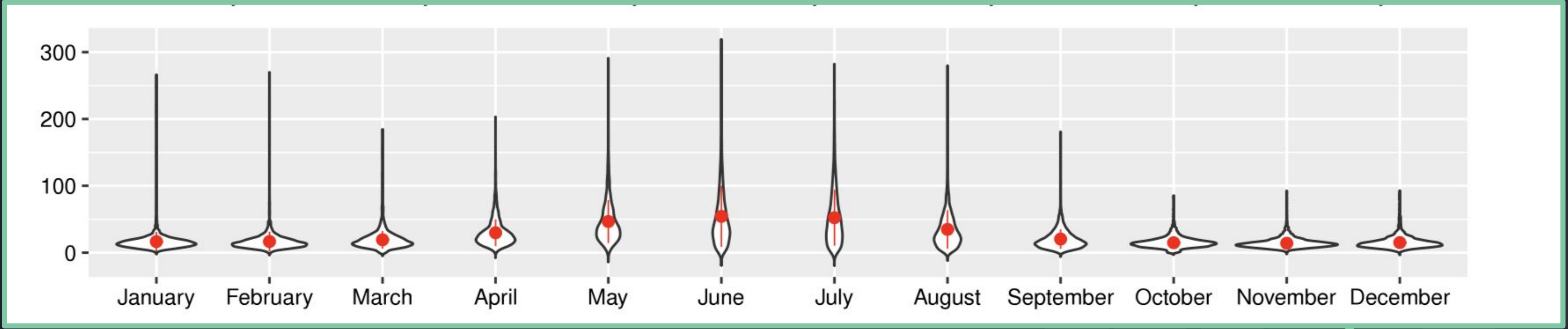


Figure 4: Monthly average, Pudahuel station January 2011 - August 2015

- April to August : more pollution; August winter months : firewood
- Idea : Temperature plays key role in pollution determinants
  - Increase in fossil fuels etc.
  - Effect on atmospheric conditions
  - What other factors can be weighed in?

Plotting : Correlations

- CO high positive correlation (0.84)
- Temperature (-0.41) & Wind Speed (-0.38) negative correlations
- Relative humidity (0.2) slightly positive correlation
- Wind Direction (-0.09) no correlation

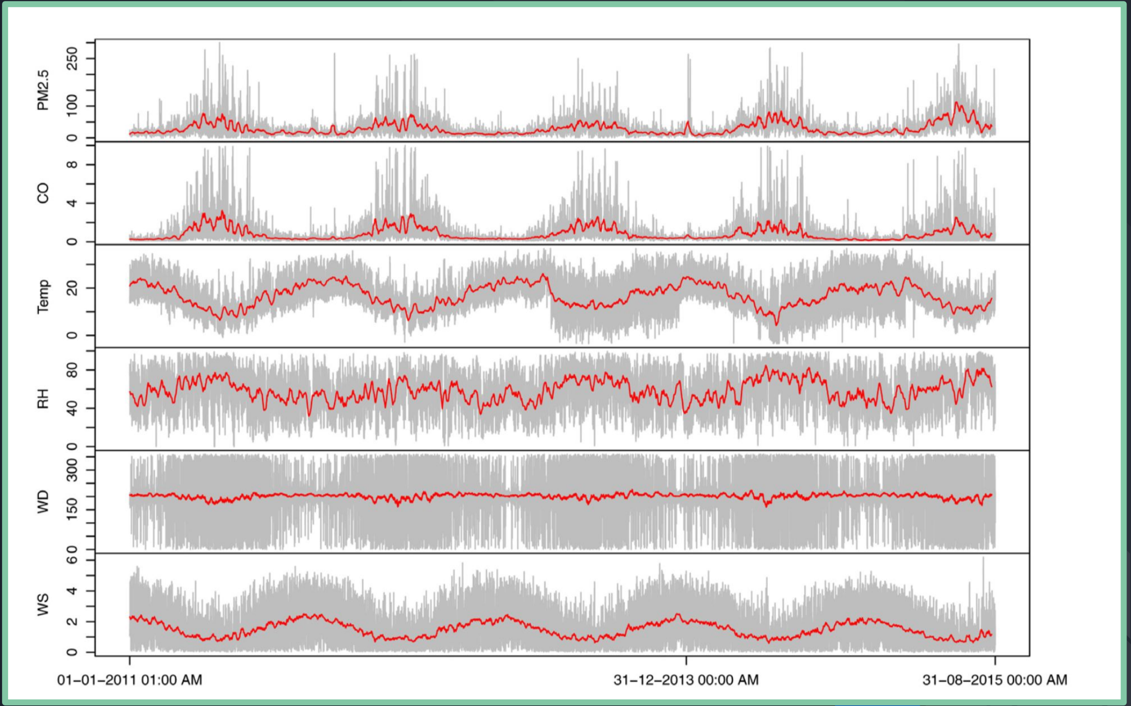
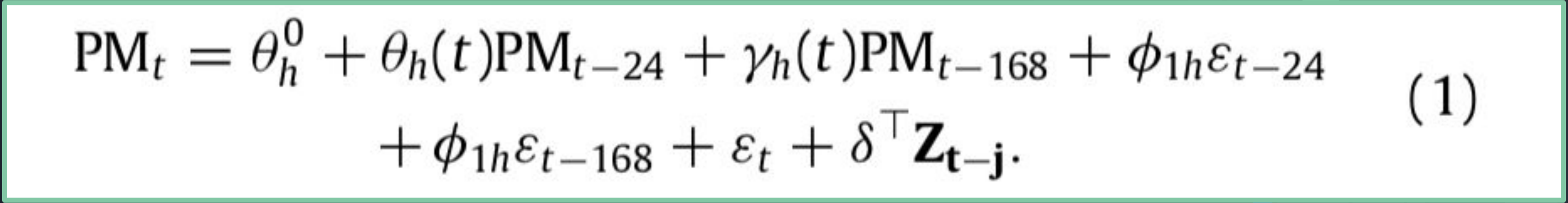


Figure 5 : hourly time series of pm2.5 concentration, CO concentration [ppb], temperature [˚C], relative humidity [%], wind direction [˚] and wind speed [m/s] for the Pudahuel monitoring station from January 2011 to August 2015.

Modelling : Model A

- Dynamic multiple equation model is proposed for PM2.5 forecasting
- 24 equations, one for each hourly interval h within a day; along with range of environmental variables

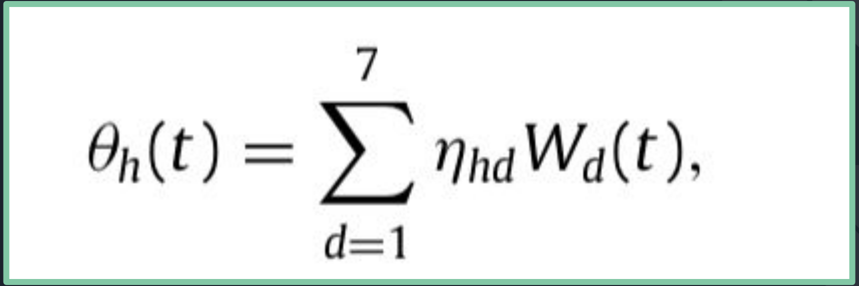


$$\begin{aligned}\mathrm{PM}_t = {} & \theta_h^0 + \theta_h(t)\mathrm{PM}_{t-24} + \gamma_h(t)\mathrm{PM}_{t-168} + \phi_{1h}\varepsilon_{t-24} \\ & + \phi_{1h}\varepsilon_{t-168} + \varepsilon_t + \delta^\top \mathbf{Z}_{\mathbf{t-j}}.\end{aligned} \tag{1}$$

- PM$_t$ denotes PM2.5 concentration at a given station at hour h = 1,2,...24 where the index t indicates pooled time series (out of 8760)

Modelling : Model A

- εt ~ N(0,σh2) is the residual term, and one-day (εt−24) and seven-day (εt−168) moving averages are included

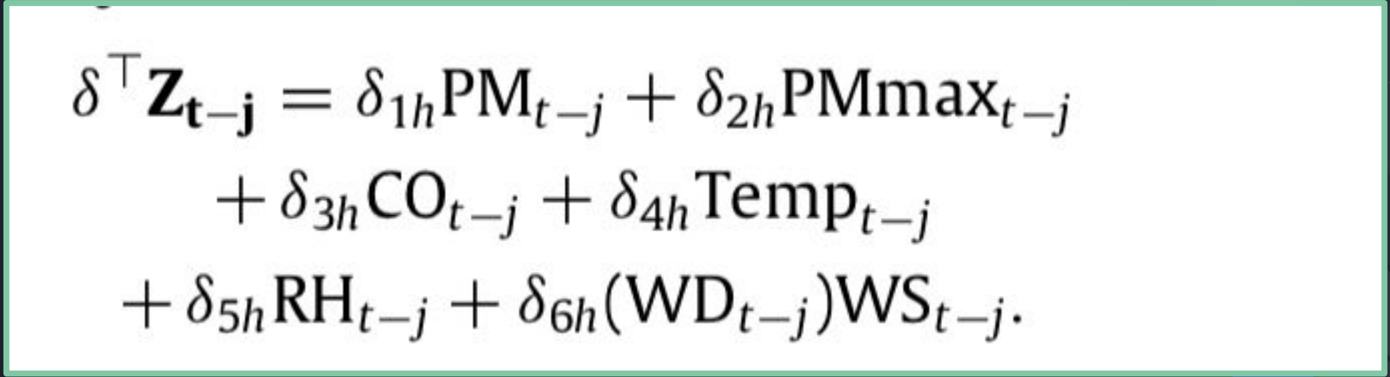


$$\theta_h(t) = \sum_{d=1}^{7} \eta_{hd} W_d(t),$$

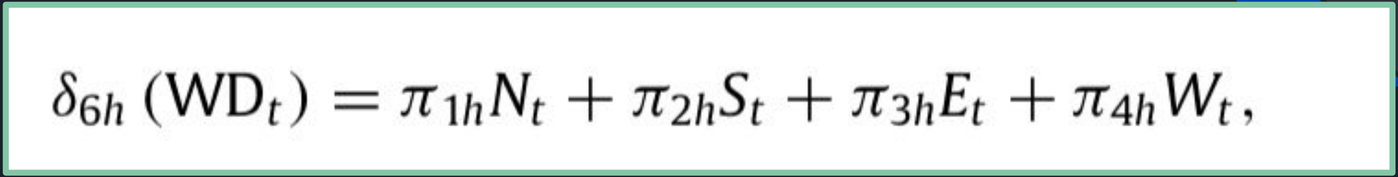
- Weekly cycle incorporated by allowing autoregressive coefficient on one-day lagged PM2.5, PM$_{t-24}$
  - W_p dummy variable - 1 if t corresponds to day of week d
  - N_{hd} coefficients to be estimated

Modelling : Model A

- As weekly pattern is included through θh(t) coefficients, similarly gamma(t) is included for annual cycle seasonality on one week lag
- Also, include set of environmental & weather covariates



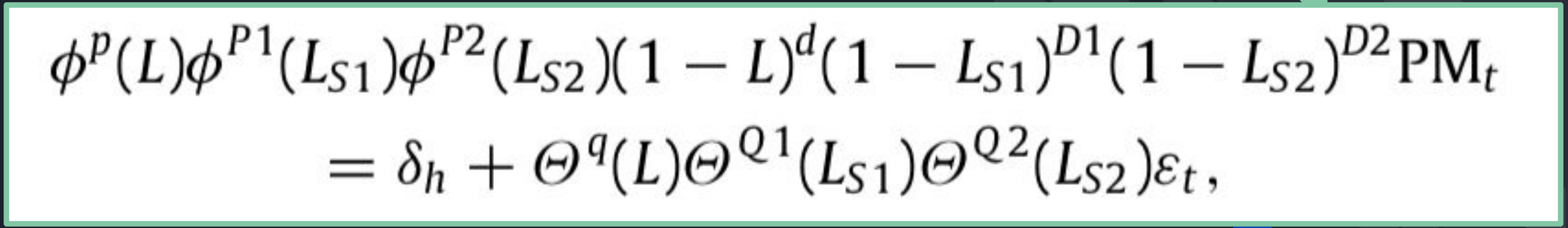
$$\begin{aligned}\delta^{\top}\mathbf{Z}_{\mathbf{t-j}} = {} & \delta_{1h}\mathrm{PM}_{t-j} + \delta_{2h}\mathrm{PMmax}_{t-j} \\ & + \delta_{3h}\mathrm{CO}_{t-j} + \delta_{4h}\mathrm{Temp}_{t-j} \\ & + \delta_{5h}\mathrm{RH}_{t-j} + \delta_{6h}(\mathrm{WD}_{t-j})\mathrm{WS}_{t-j}.\end{aligned}$$



$$\delta_{6h}\left(\mathrm{WD}_t\right) = \pi_{1h}N_t + \pi_{2h}S_t + \pi_{3h}E_t + \pi_{4h}W_t\,,$$

Comparison : Model B

- SARIMAX model :
  - L delay operator
  - Phi - autoregressive polynomial of order p
  - Theta - moving average of order q

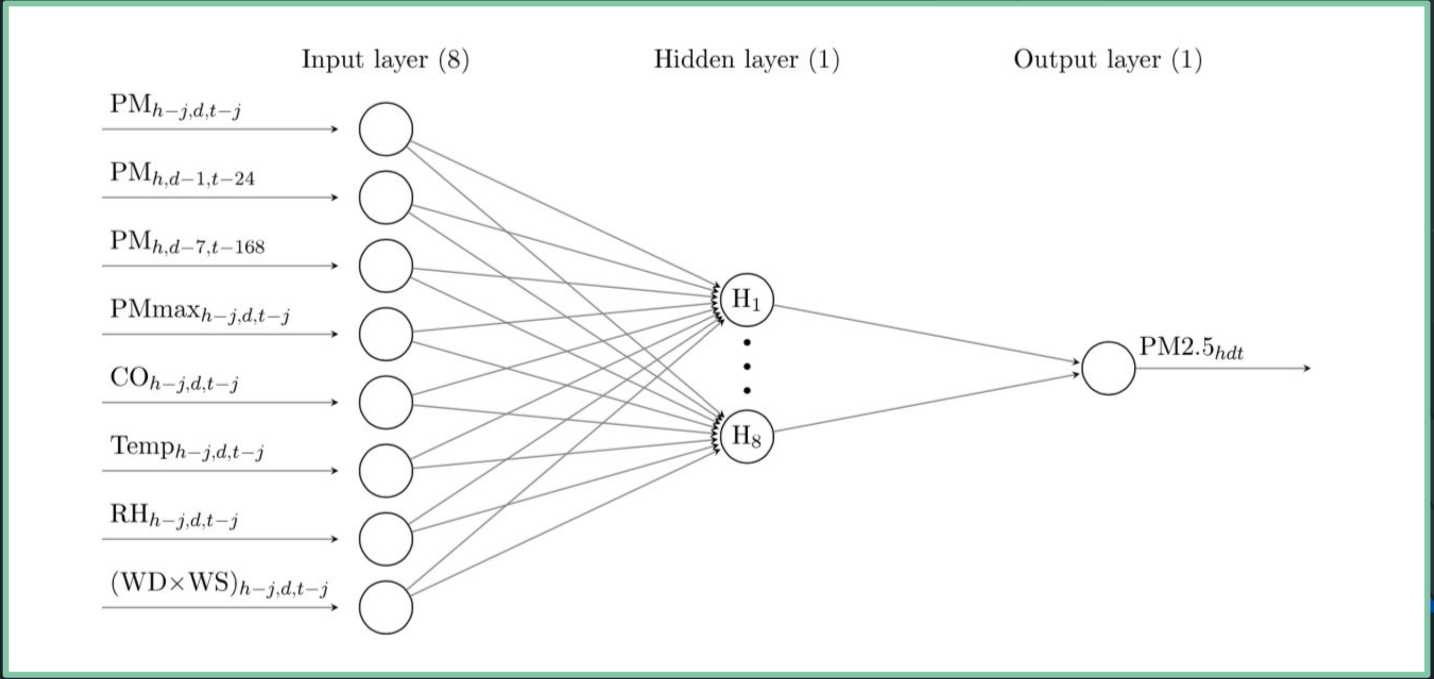


$$\phi^{p}(L)\phi^{P1}(L_{S1})\phi^{P2}(L_{S2})(1-L)^{d}(1-L_{S1})^{D1}(1-L_{S2})^{D2}\mathrm{PM}_t \\ = \delta_h + \Theta^{q}(L)\Theta^{Q1}(L_{S1})\Theta^{Q2}(L_{S2})\varepsilon_t ,$$

- DME is SARIMAX $(1, 0, 1) \times (1, 0, 1)_{24} \times (1, 0, 0)_{168}$

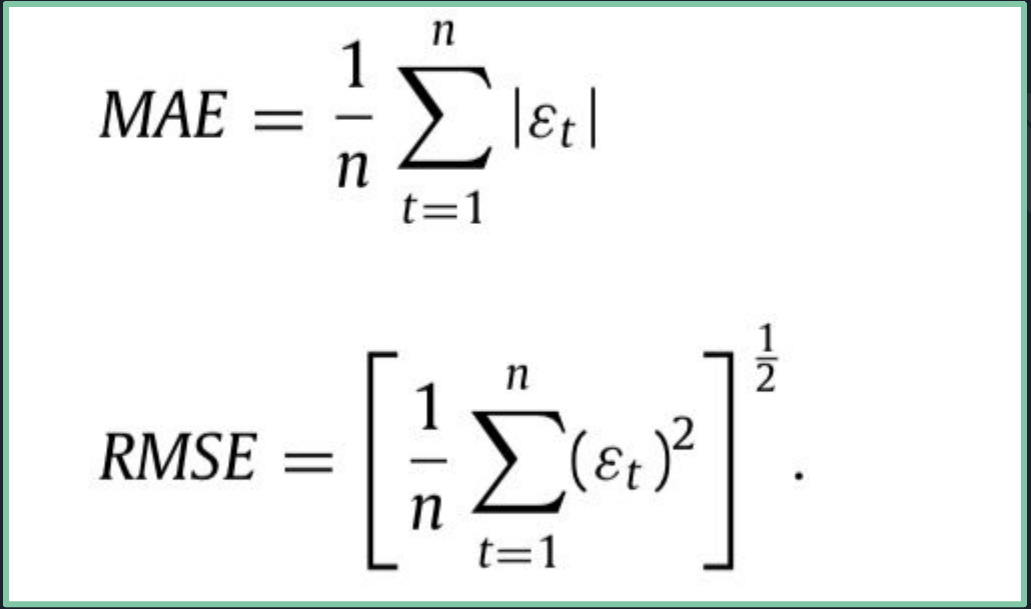
Comparison : Model C

- ANN Model


Input layer (8)
Hidden layer (1)
Output layer (1)
$\mathrm{PM}_{h-j,d,t-j}$
$\mathrm{PM}_{h,d-1,t-24}$
$\mathrm{PM}_{h,d-7,t-168}$
$\mathrm{PMmax}_{h-j,d,t-j}$
$\mathrm{CO}_{h-j,d,t-j}$
$\mathrm{Temp}_{h-j,d,t-j}$
$\mathrm{RH}_{h-j,d,t-j}$
$(\mathrm{WD}\times\mathrm{WS})_{h-j,d,t-j}$
H_1
H_8
$\mathrm{PM2.5}_{hdt}$

Metrics

- Mean Absolute Error and Root Mean Square Error
  - Evaluating fit for forecast for time series
  - E_t : Observed - Forecasted value



$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\varepsilon_t|$$

$$RMSE = \left[\frac{1}{n}\sum_{t=1}^{n}(\varepsilon_t)^2\right]^{\frac{1}{2}}.$$

Results

- DME vs SARIMAX, ANN
- Three periods of estimation : 2011-13; 2012-14; 2013-15
- 1 hr and 24 hr forecast

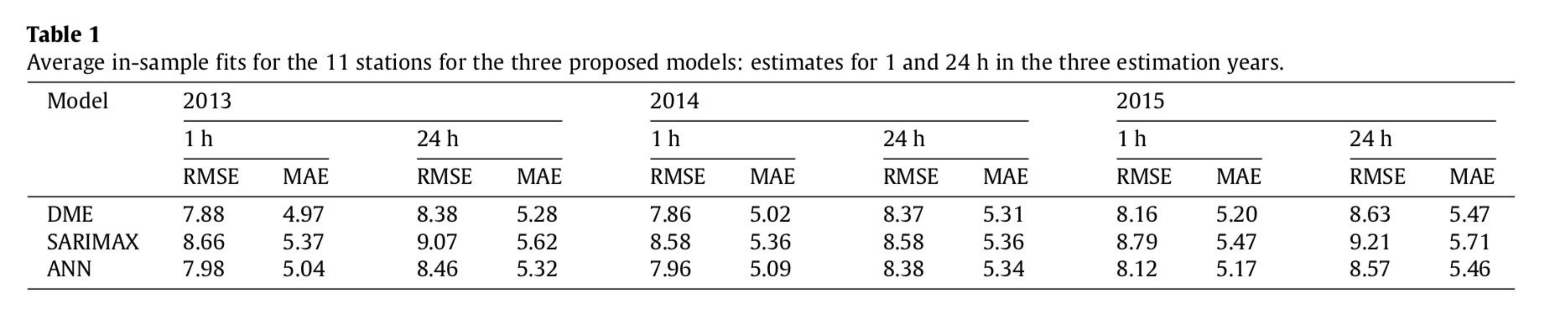
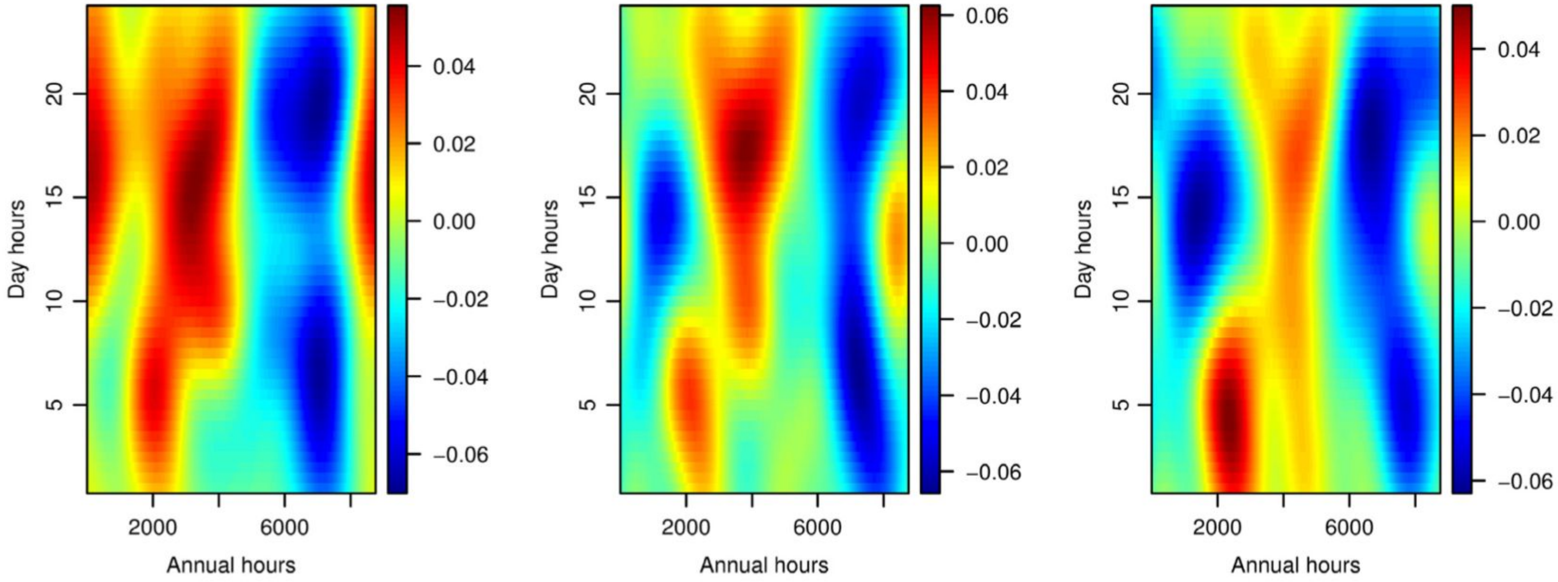


Table 1
Average in-sample fits for the 11 stations for the three proposed models: estimates for 1 and 24 h in the three estimation years.

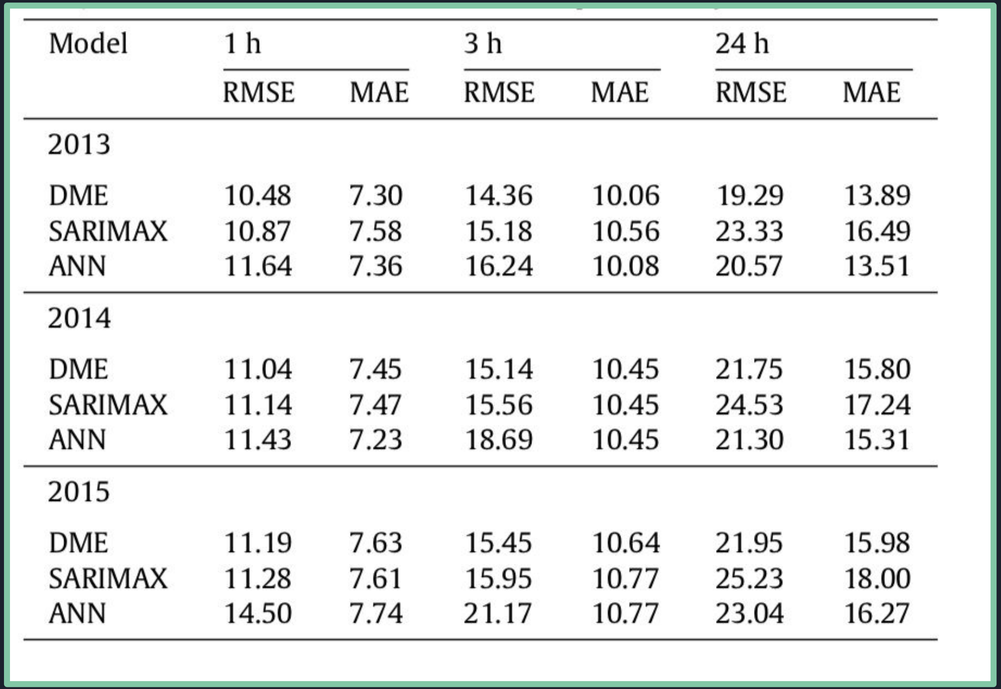
Model	2013				2014				2015			
	1 h		24 h		1 h		24 h		1 h		24 h	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
DME	7.88	4.97	8.38	5.28	7.86	5.02	8.37	5.31	8.16	5.20	8.63	5.47
SARIMAX	8.66	5.37	9.07	5.62	8.58	5.36	8.58	5.36	8.79	5.47	9.21	5.71
ANN	7.98	5.04	8.46	5.32	7.96	5.09	8.38	5.34	8.12	5.17	8.57	5.46

Results



Blue implies lower persistence for week, while red reflects high persistence

Results



Model	1 h		3 h		24 h	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
2013						
DME	10.48	7.30	14.36	10.06	19.29	13.89
SARIMAX	10.87	7.58	15.18	10.56	23.33	16.49
ANN	11.64	7.36	16.24	10.08	20.57	13.51
2014						
DME	11.04	7.45	15.14	10.45	21.75	15.80
SARIMAX	11.14	7.47	15.56	10.45	24.53	17.24
ANN	11.43	7.23	18.69	10.45	21.30	15.31
2015						
DME	11.19	7.63	15.45	10.64	21.95	15.98
SARIMAX	11.28	7.61	15.95	10.77	25.23	18.00
ANN	14.50	7.74	21.17	10.77	23.04	16.27

Results

- 76% success rate for alert episodes; 100% for emergency episodes, 2013

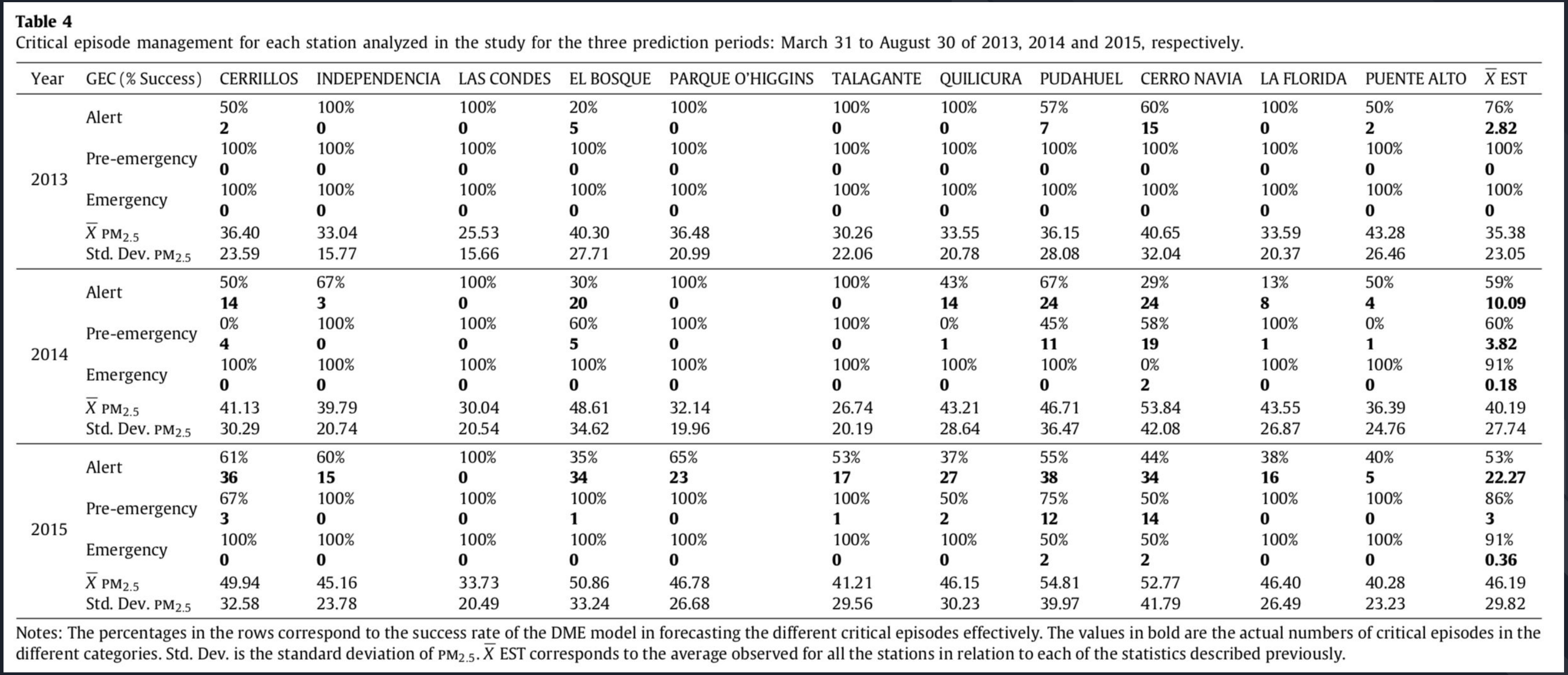


Table 4
Critical episode management for each station analyzed in the study for the three prediction periods: March 31 to August 30 of 2013, 2014 and 2015, respectively.

Year	GEC (% Success)	CERRILLOS	INDEPENDENCIA	LAS CONDES	EL BOSQUE	PARQUE O'HIGGINS	TALAGANTE	QUILICURA	PUDAHUEL	CERRO NAVIA	LA FLORIDA	PUENTE ALTO	$\bar{X}$ EST
2013	Alert	50% **2**	100% **0**	100% **0**	20% **5**	100% **0**	100% **0**	100% **0**	57% **7**	60% **15**	100% **0**	50% **2**	76% **2.82**
	Pre-emergency	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**
	Emergency	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**
	$\bar{X}$ PM$_{2.5}$	36.40	33.04	25.53	40.30	36.48	30.26	33.55	36.15	40.65	33.59	43.28	35.38
	Std. Dev. PM$_{2.5}$	23.59	15.77	15.66	27.71	20.99	22.06	20.78	28.08	32.04	20.37	26.46	23.05
2014	Alert	50% **14**	67% **3**	100% **0**	30% **20**	100% **0**	100% **0**	43% **14**	67% **24**	29% **24**	13% **8**	50% **4**	59% **10.09**
	Pre-emergency	0% **4**	100% **0**	100% **0**	60% **5**	100% **0**	100% **0**	0% **1**	45% **11**	58% **19**	100% **1**	0% **1**	60% **3.82**
	Emergency	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	0% **2**	100% **0**	100% **0**	91% **0.18**
	$\bar{X}$ PM$_{2.5}$	41.13	39.79	30.04	48.61	32.14	26.74	43.21	46.71	53.84	43.55	36.39	40.19
	Std. Dev. PM$_{2.5}$	30.29	20.74	20.54	34.62	19.96	20.19	28.64	36.47	42.08	26.87	24.76	27.74
2015	Alert	61% **36**	60% **15**	100% **0**	35% **34**	65% **23**	53% **17**	37% **27**	55% **38**	44% **34**	38% **16**	40% **5**	53% **22.27**
	Pre-emergency	67% **3**	100% **0**	100% **0**	100% **1**	100% **0**	100% **1**	50% **2**	75% **12**	50% **14**	100% **0**	100% **0**	86% **3**
	Emergency	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	100% **0**	50% **2**	50% **2**	100% **0**	100% **0**	91% **0.36**
	$\bar{X}$ PM$_{2.5}$	49.94	45.16	33.73	50.86	46.78	41.21	46.15	54.81	52.77	46.40	40.28	46.19
	Std. Dev. PM$_{2.5}$	32.58	23.78	20.49	33.24	26.68	29.56	30.23	39.97	41.79	26.49	23.23	29.82

Notes: The percentages in the rows correspond to the success rate of the DME model in forecasting the different critical episodes effectively. The values in bold are the actual numbers of critical episodes in the different categories. Std. Dev. is the standard deviation of PM$_{2.5}$. $\bar{X}$ EST corresponds to the average observed for all the stations in relation to each of the statistics described previously.

Interpretation

- Less susceptible to overfitting like ANN
- Linear hence takes less time and resources
- DME can capture seasonality adequately and surpasses competing ANN, SARIMAX
- Can be used for prediction by government institutions for PM 2.5 predictions
- Extension : New Delhi, India

Thank You!