Phân tích ô nhiễm không khí ở VN

Sử dụng R & Openair package

Tuan Vu
Senior air quality scientist, King's College London
Email: tuan.vu@kcl.ac.uk
https://tuanvvu.github.io/

10-April-2020

Aims

- Introduction to R data analysis software
- Introduction and use of the R package: ***openair***
- Machine learning in data analysis

“We can only see a short distance ahead, but we can see plenty there that needs to be done”- Alan Turning

I. Introduction to R data analysis software

1. Downloading and installing R/ Rstudio

2. General approach to data analysis

- Use scripts: save all objects in the current R sessions as an .RData file
- Leave the data alone: as much as you can
- Coding style
- Simple R and vectors: R cheat sheet

2. Useful packages

- lubridate/dplyr/plyr
- ggplot2
- openair/worldmet

II. Introduction to *"openair"*

http://www.openair-project.org/PDF/OpenAir_Manual.pdf

Useful *openair* functions

1. Summary data: Understand your data

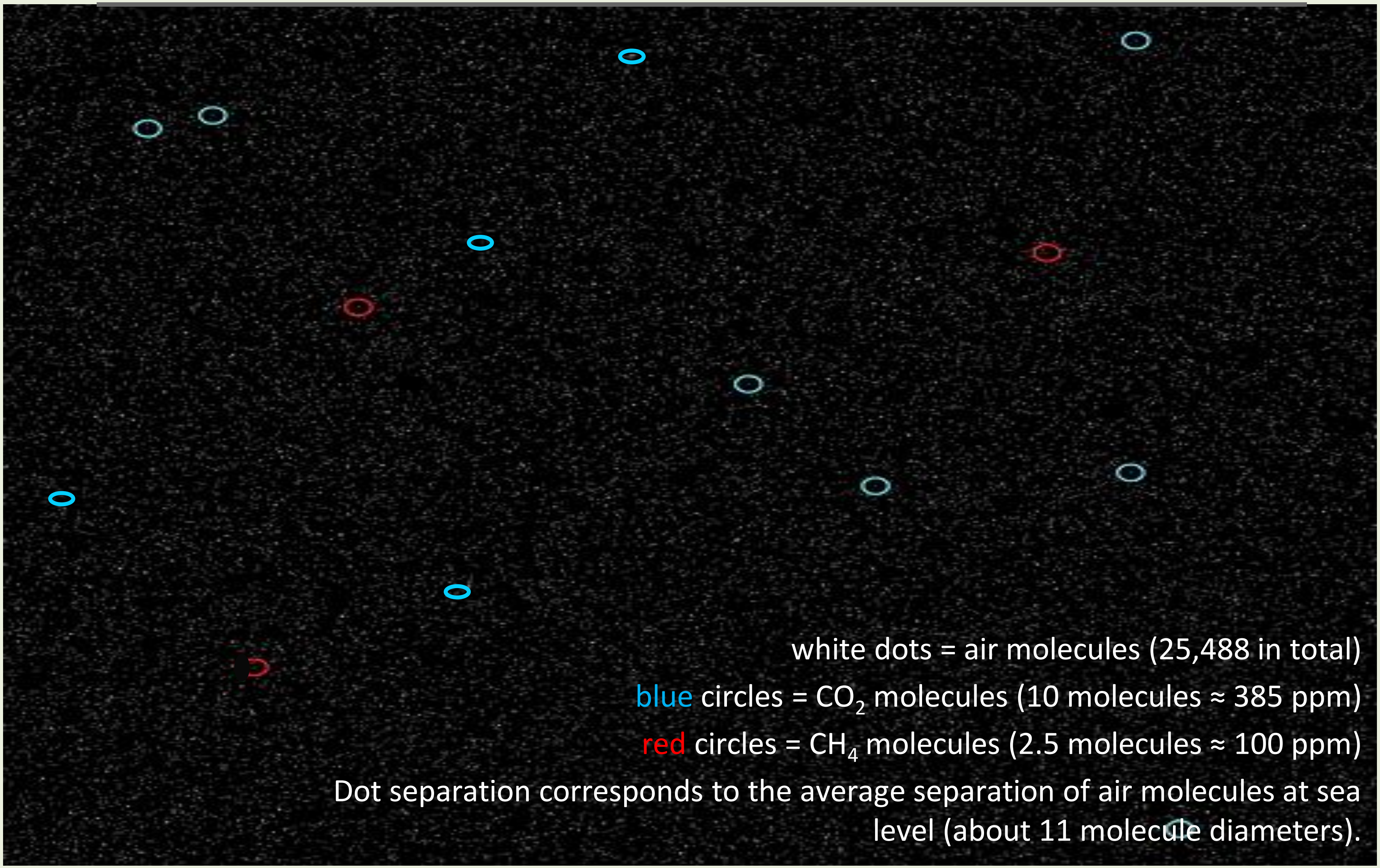
2. Merging data sets

3. Selecting data by date

4. Averaging data to different time intervals

5. The *ScatterPlot*

Understanding about the pollutants



white dots = air molecules (25,488 in total)

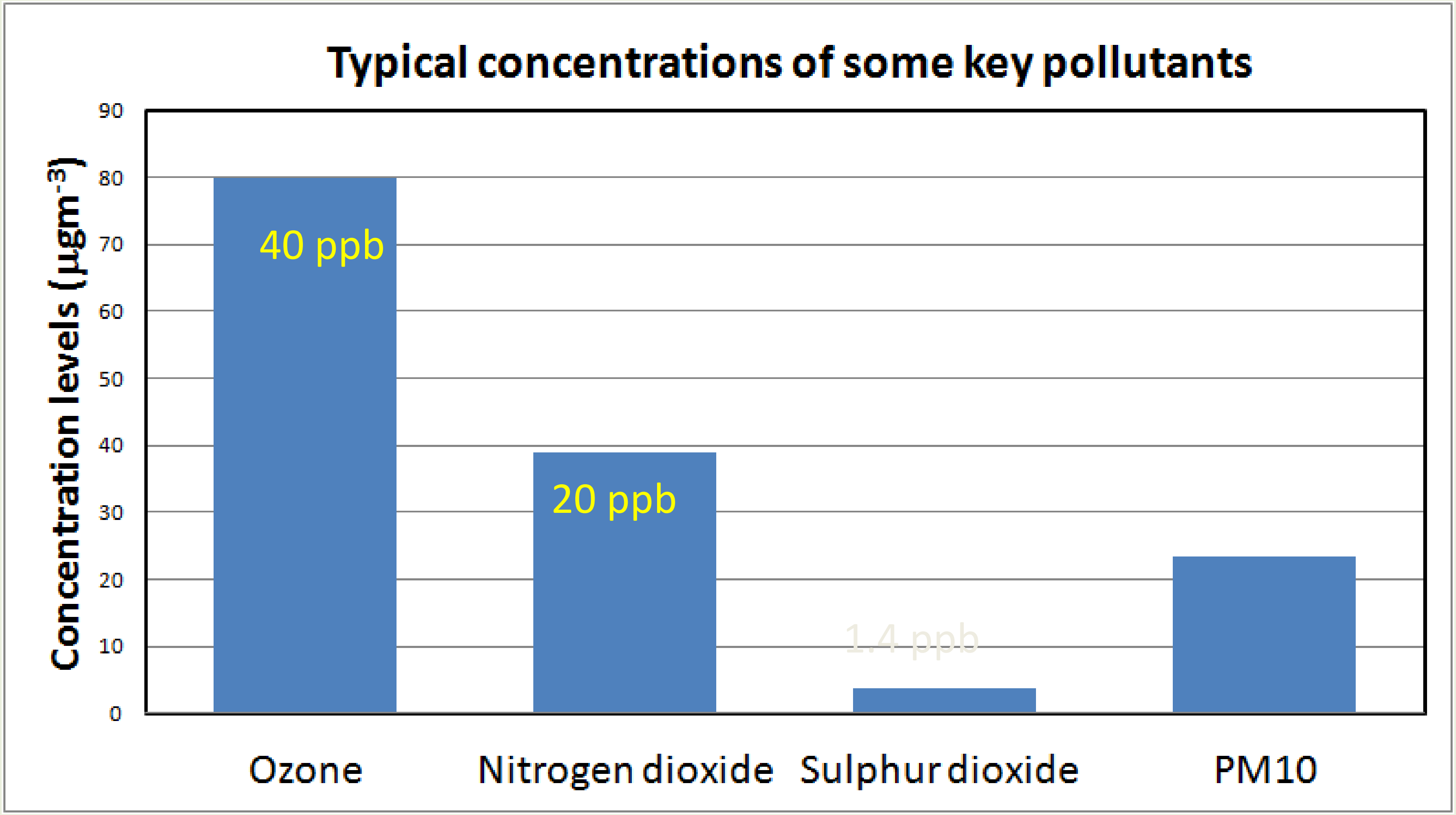
blue circles = CO_2 molecules (10 molecules ≈ 385 ppm)

red circles = CH_4 molecules (2.5 molecules ≈ 100 ppm)

Dot separation corresponds to the average separation of air molecules at sea level (about 11 molecule diameters).

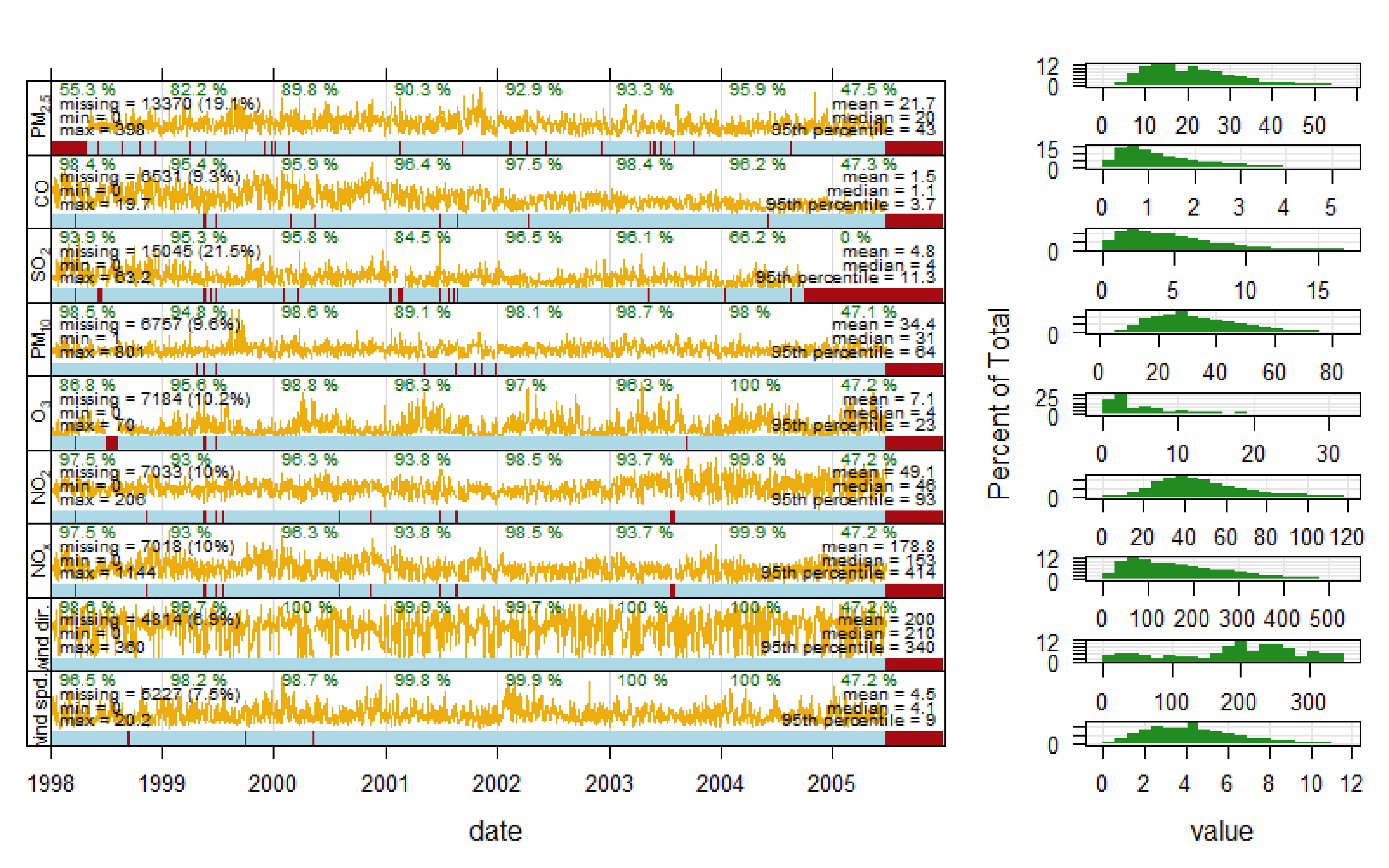
Link to rice comparison

Levels of Pollutants

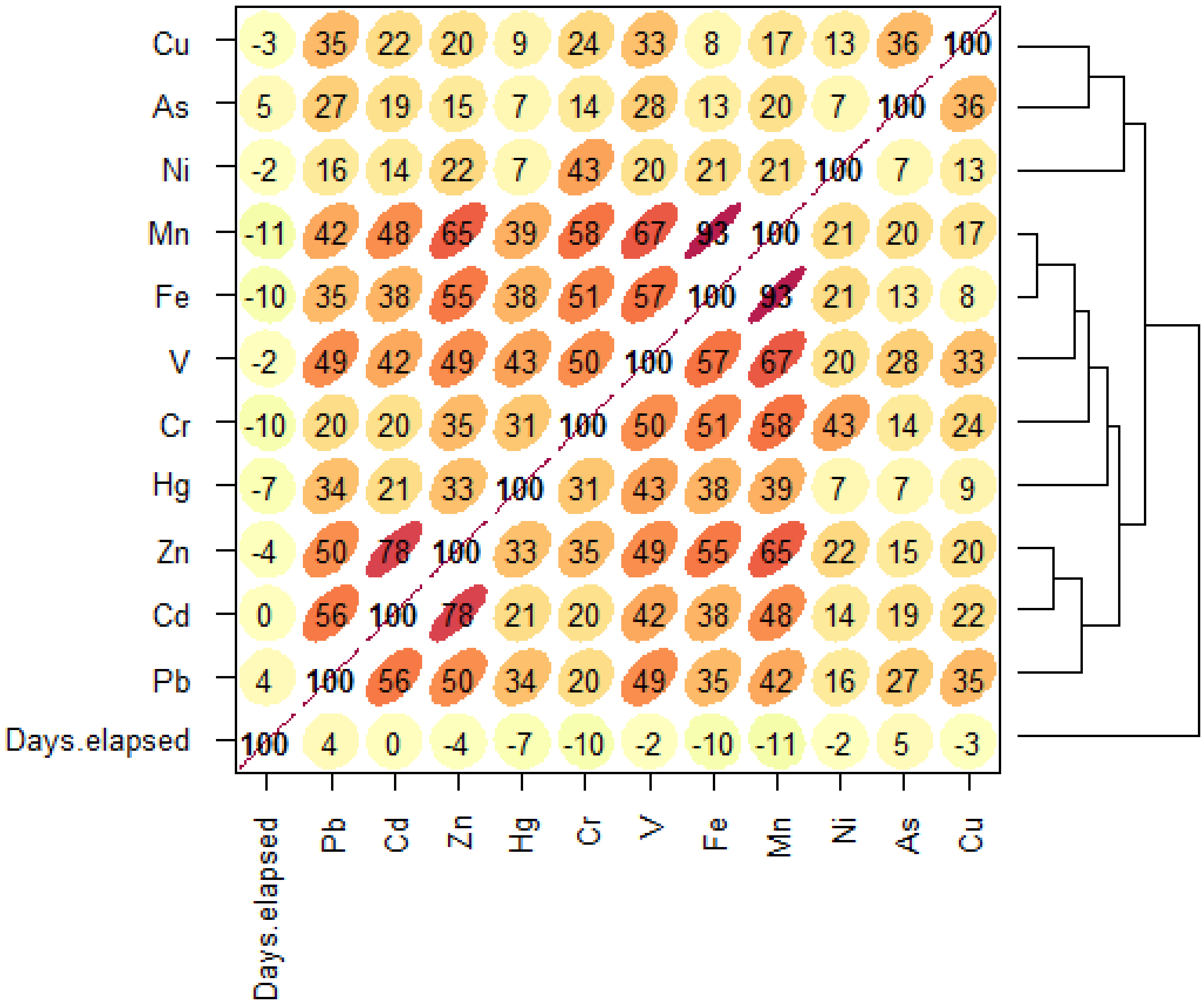


Data obtained from: http://uk-air.defra.gov.uk/

summaryPlot

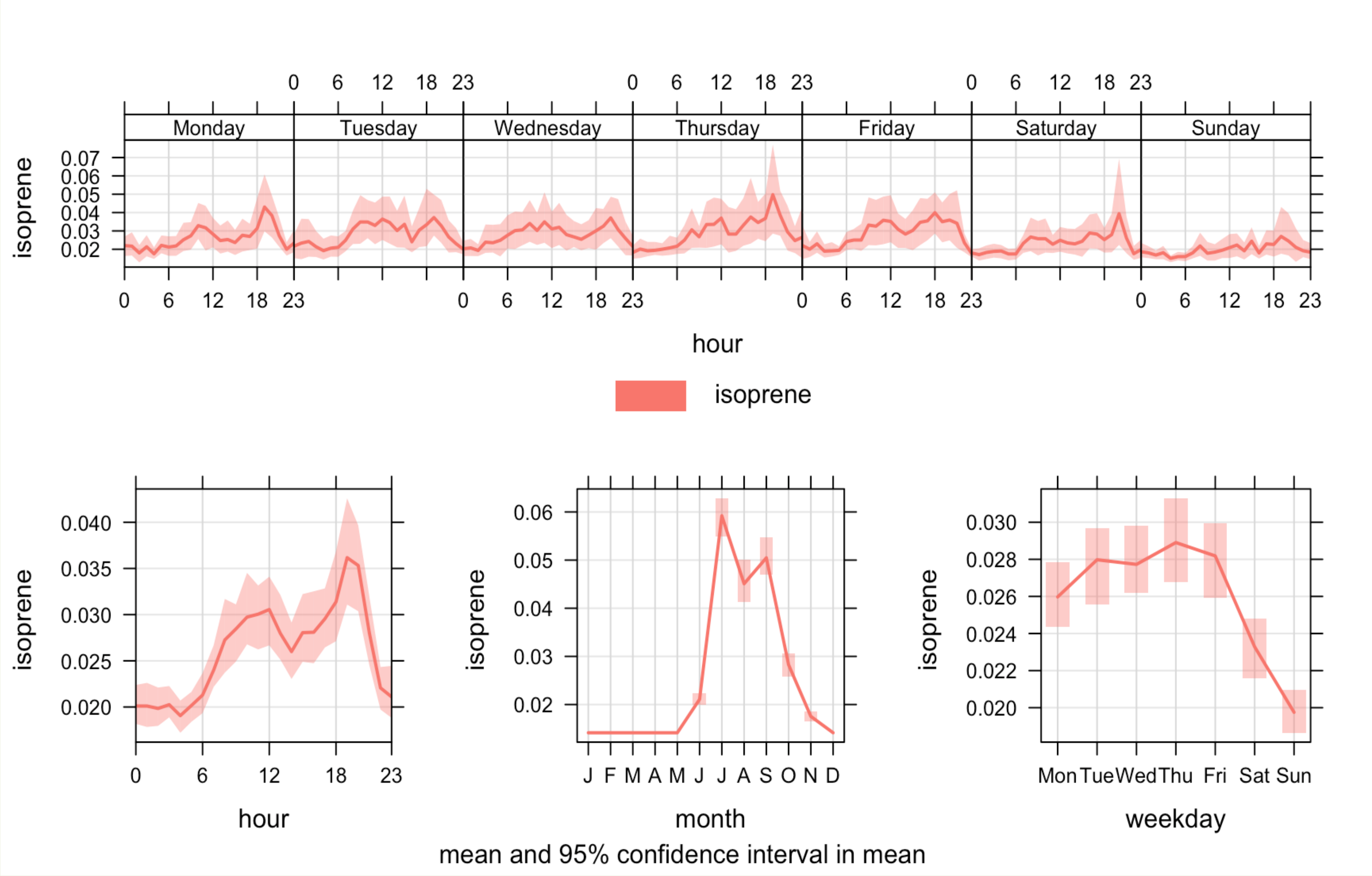


6. CorPlot



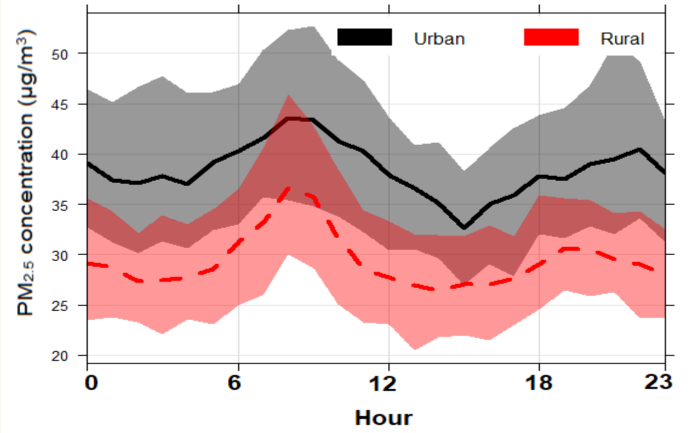
	Days.elapsed	Pb	Cd	Zn	Hg	Cr	V	Fe	Mn	Ni	As	Cu
Cu	-3	35	22	20	9	24	33	8	17	13	36	100
As	5	27	19	15	7	14	28	13	20	7	100	36
Ni	-2	16	14	22	7	43	20	21	21	100	7	13
Mn	-11	42	48	65	39	58	67	93	100	21	20	17
Fe	-10	35	38	55	38	51	57	100	93	21	13	8
V	-2	49	42	49	43	50	100	57	67	20	28	33
Cr	-10	20	20	35	31	100	50	51	58	43	14	24
Hg	-7	34	21	33	100	31	43	38	39	7	7	9
Zn	-4	50	78	100	33	35	49	55	65	22	15	20
Cd	0	56	100	78	21	20	42	38	48	14	19	22
Pb	4	100	56	50	34	20	49	35	42	16	27	35
Days.elapsed	100	4	0	-4	-7	-10	-2	-10	-11	-2	5	-3

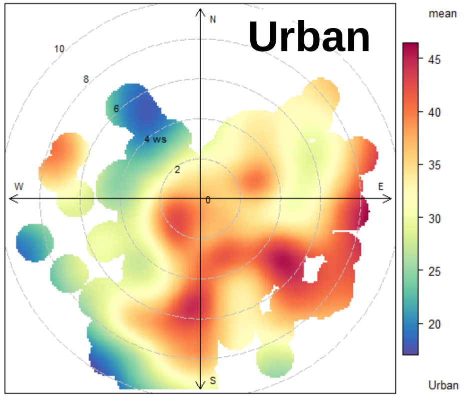
6. Time Variable Function

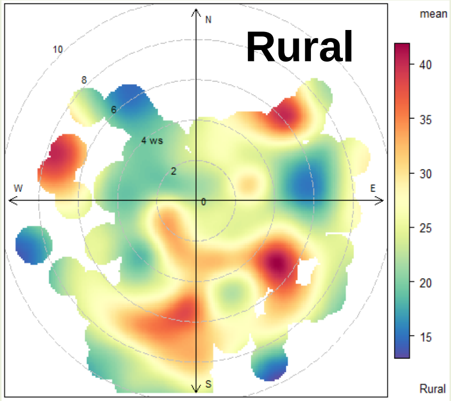


PM$_{2.5}$ diurnal patterns & Polar Plot

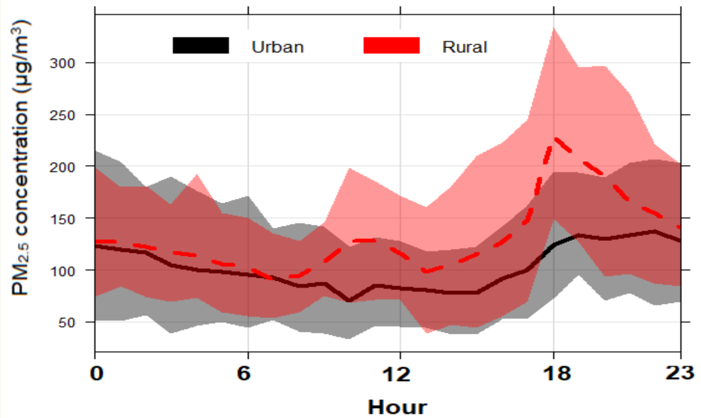
Summer 2017

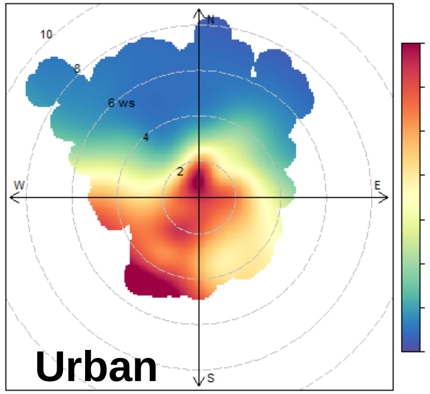


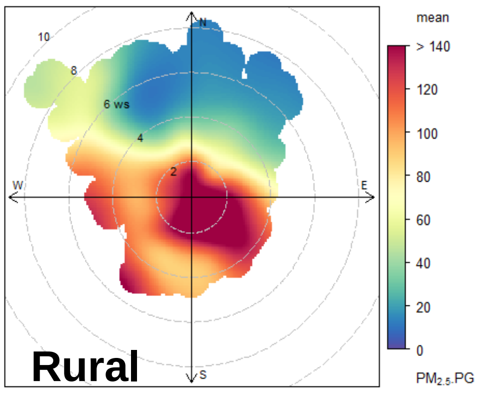




Winter 2016

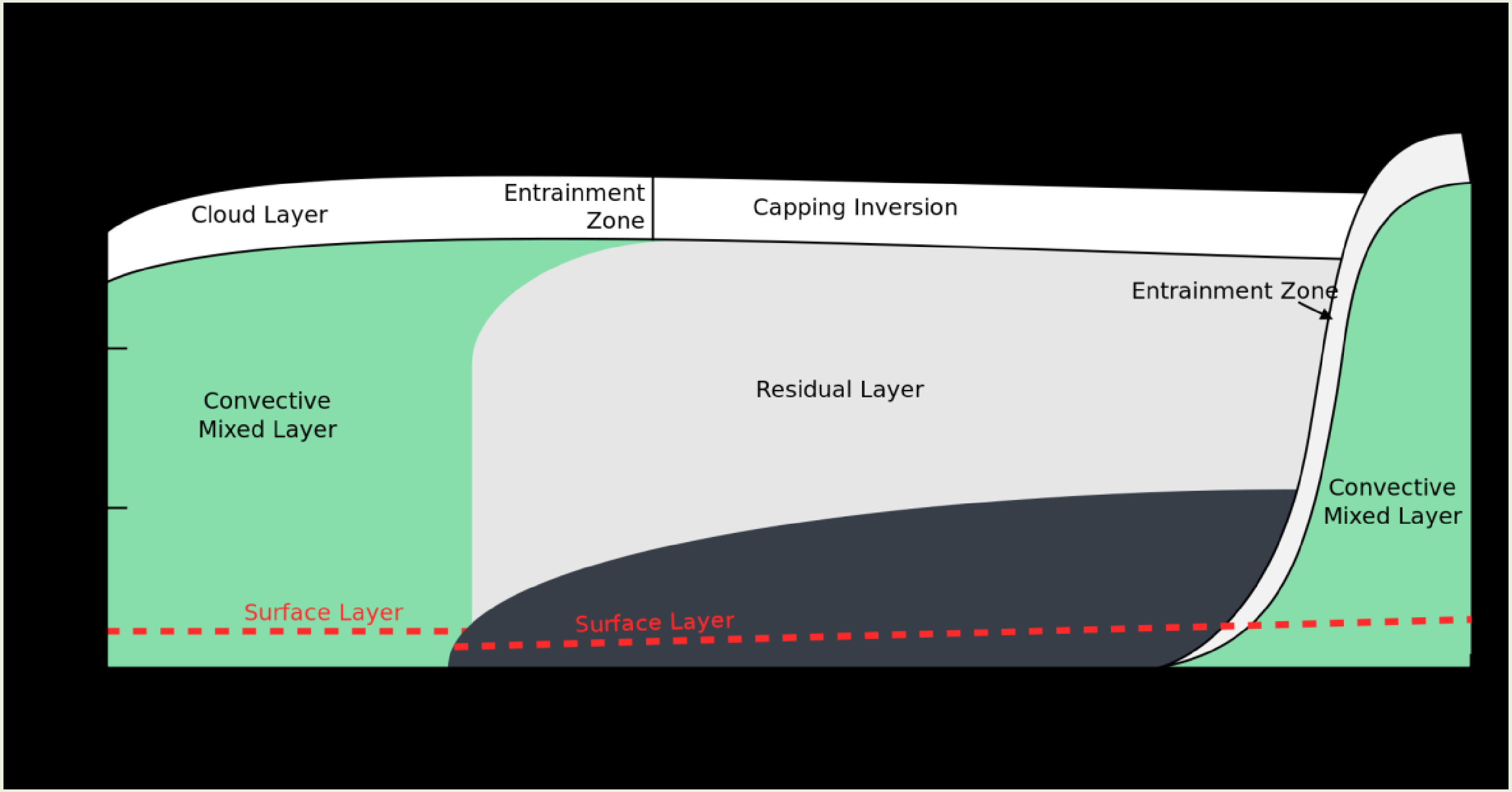






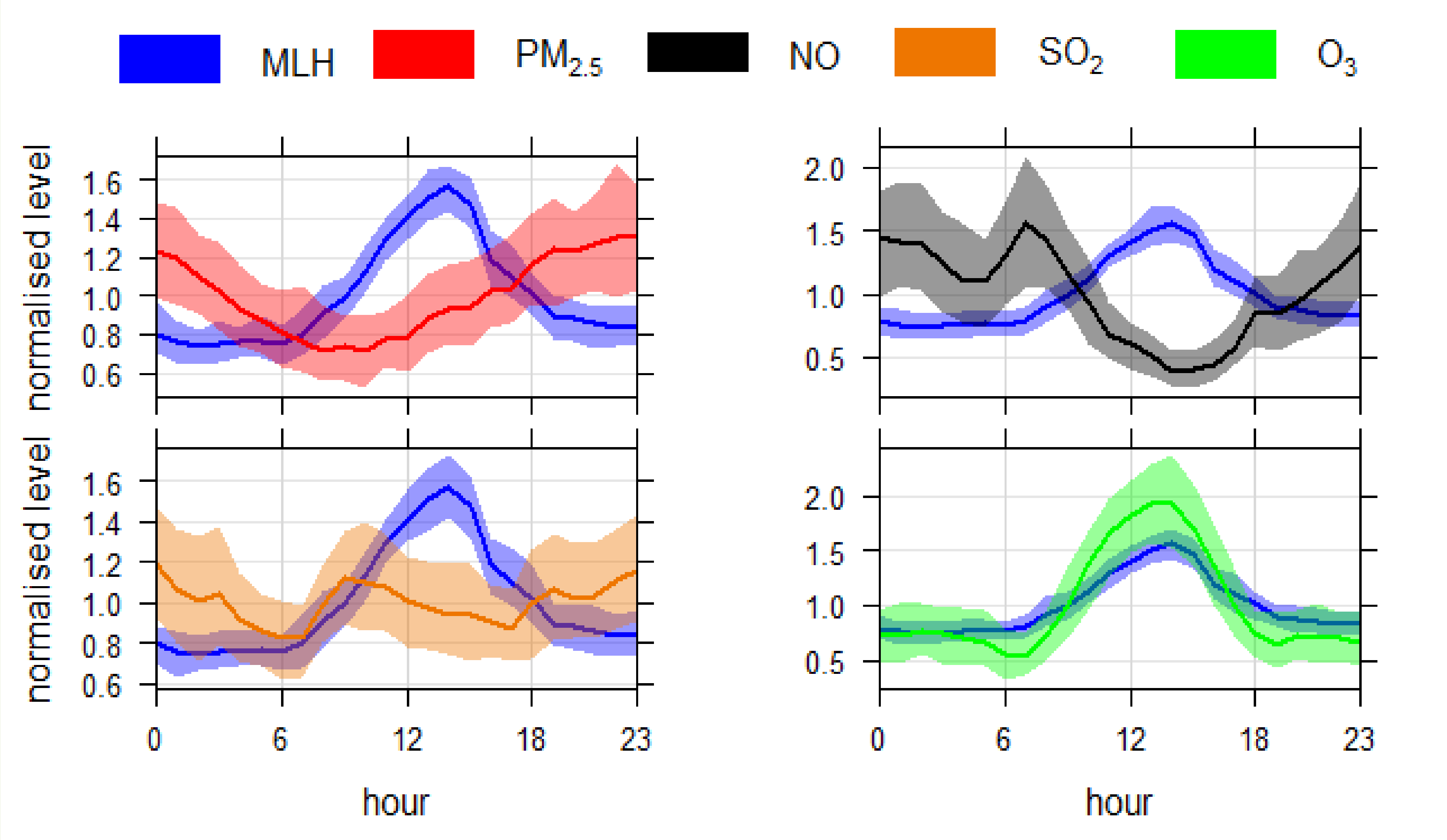
Hourly PM2.5 data in Pinggu was provided by PKU

Diurnal variations in the ABL

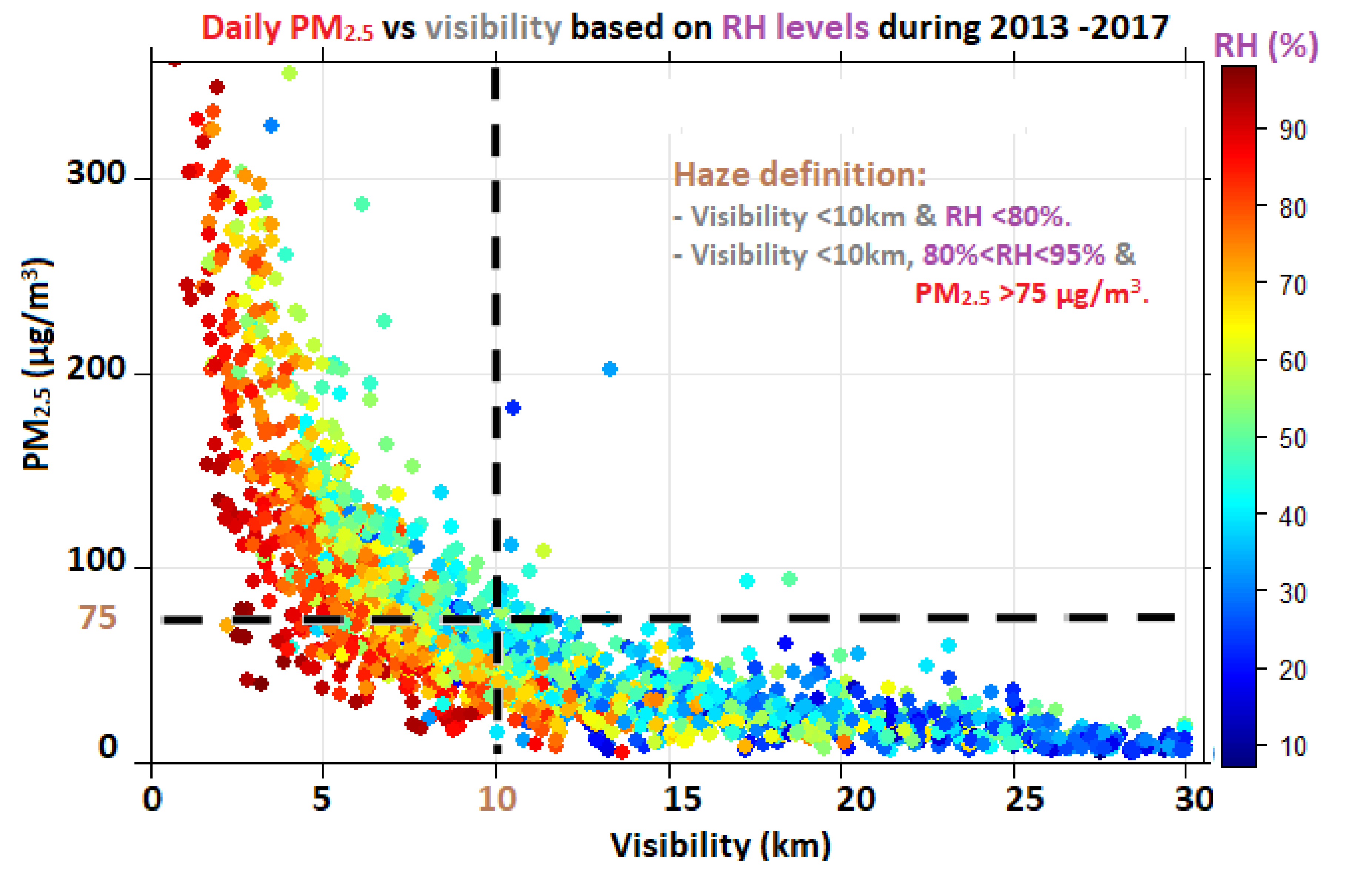


Cooling of surface during sunset – stable nocturnal layer grows from below

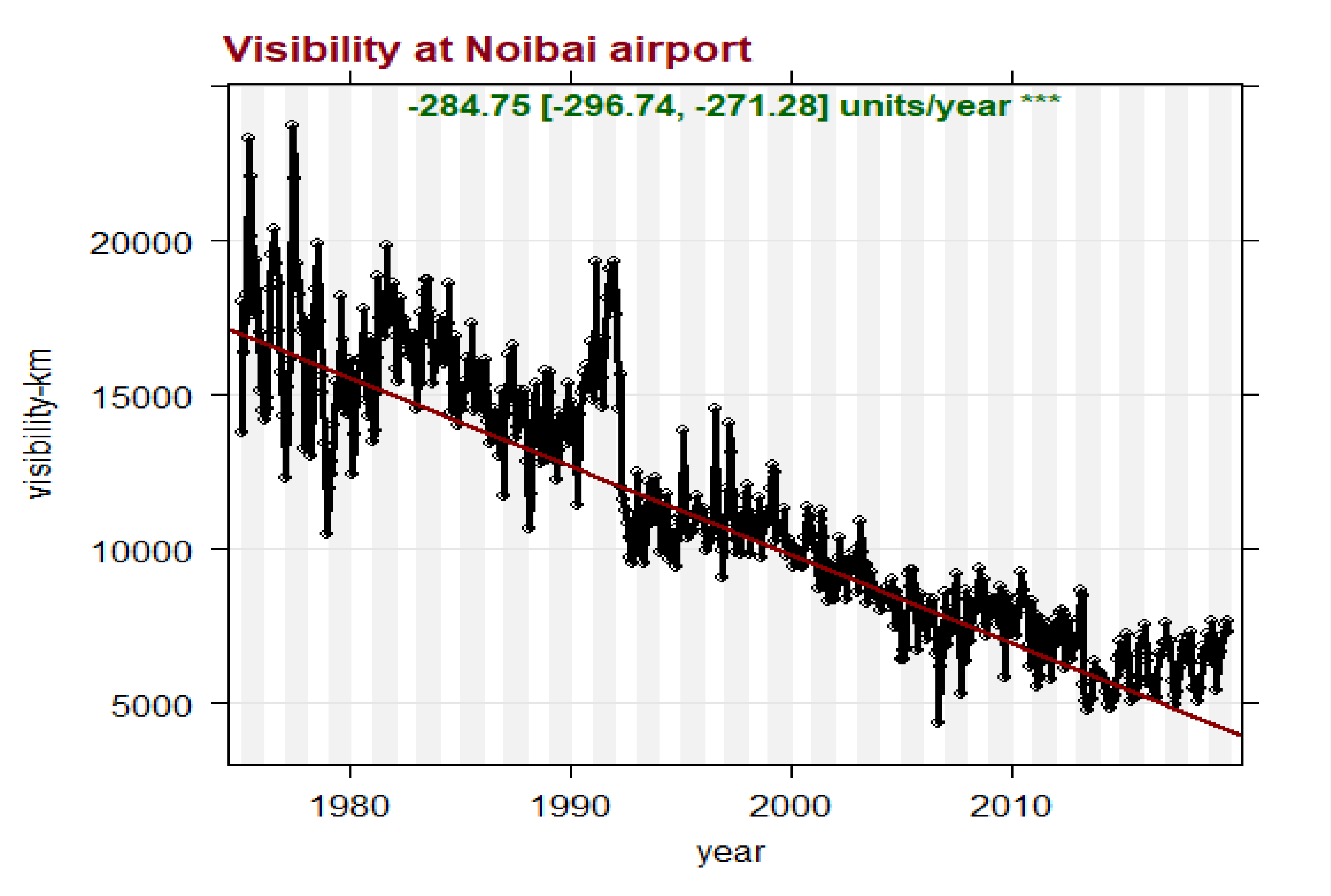
Heating from radiation on surface during sunrise - convection breaks up stable nocturnal later and entrains air from above


MLH
PM2.5
NO
SO2
O3
normalised level
1.6
1.4
1.2
1.0
0.8
0.6
2.0
1.5
1.0
0.5
0
6
12
18
23
hour

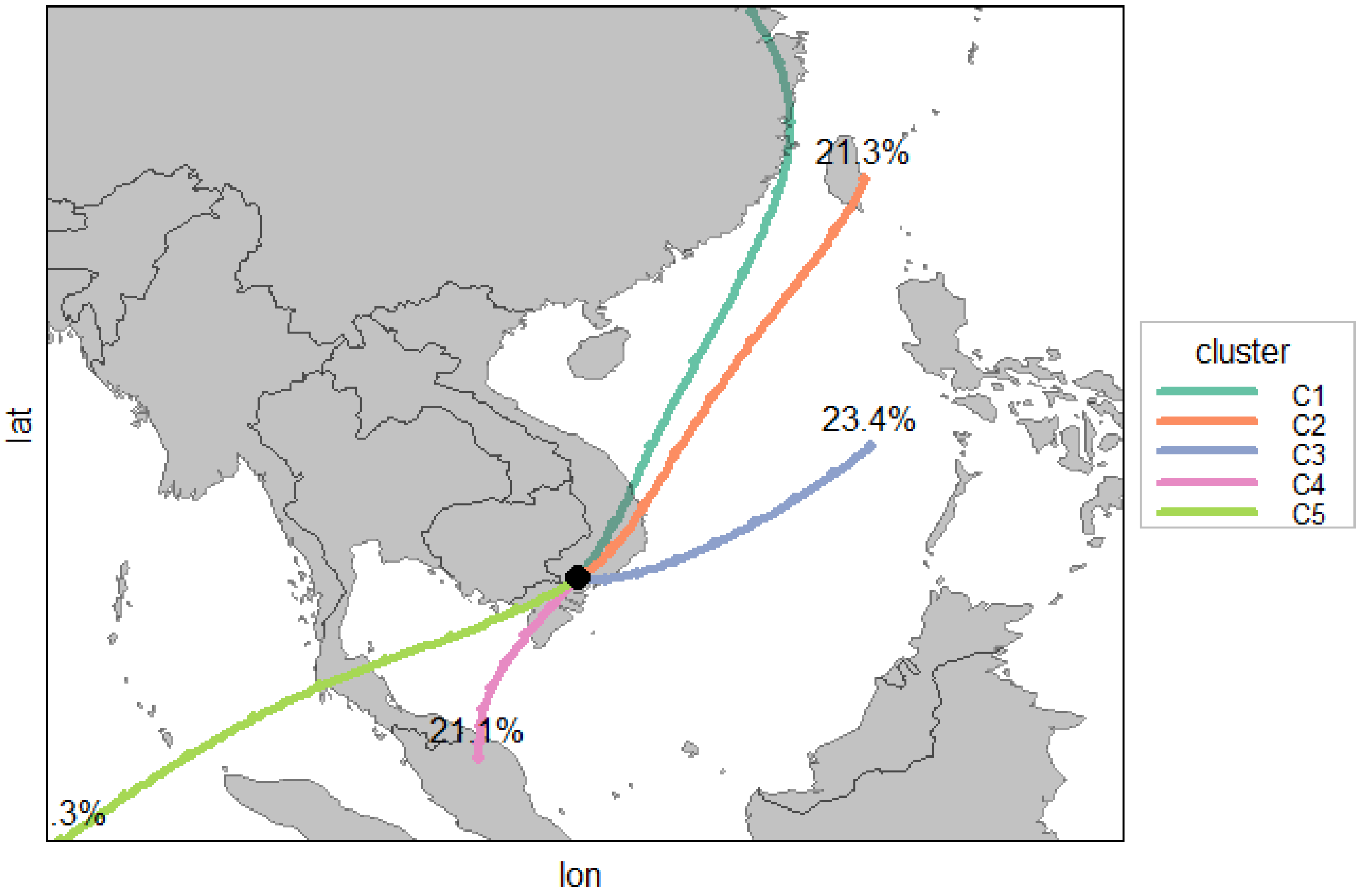
7. Scatter plot



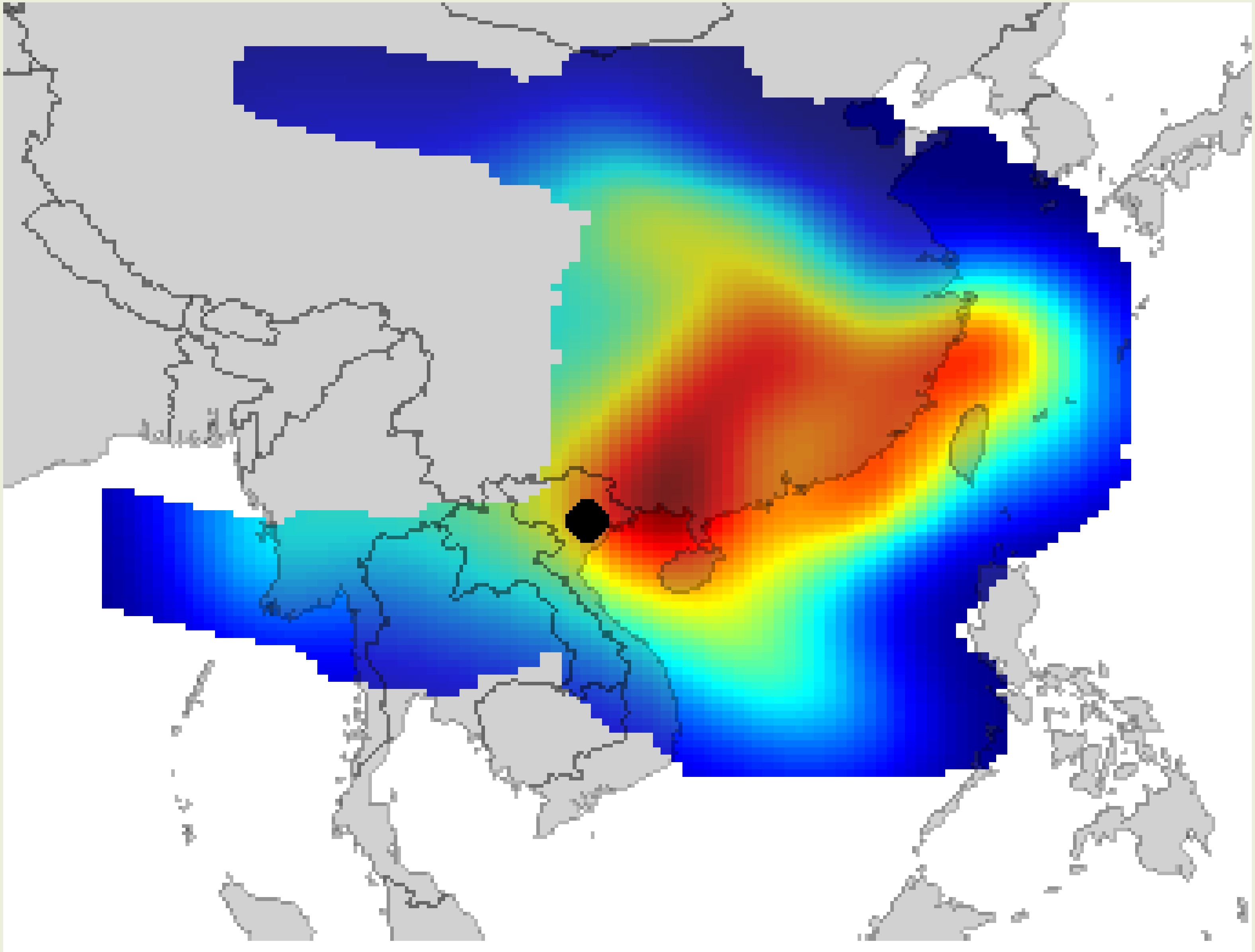
8. Theilsen regression



9. Air-cluster analysis



10. CTW longrange/transport



III. Machine learning

Useful *technique*

1. Factor analysis: PMF, K-mean cluster
2. Decision tree: Random forest, BRT
3. Deep learning: CCN

https://machinelearningcoban.com/
https://rpubs.com/lengockhanhi

Other program: Python

Weather normalisation using package “*rmweather*”

https://github.com/skgrange/rmweather

Random forest algorithm:

- What is decision tree?
- Random forest is non-linear regression?
- How to select the trees?

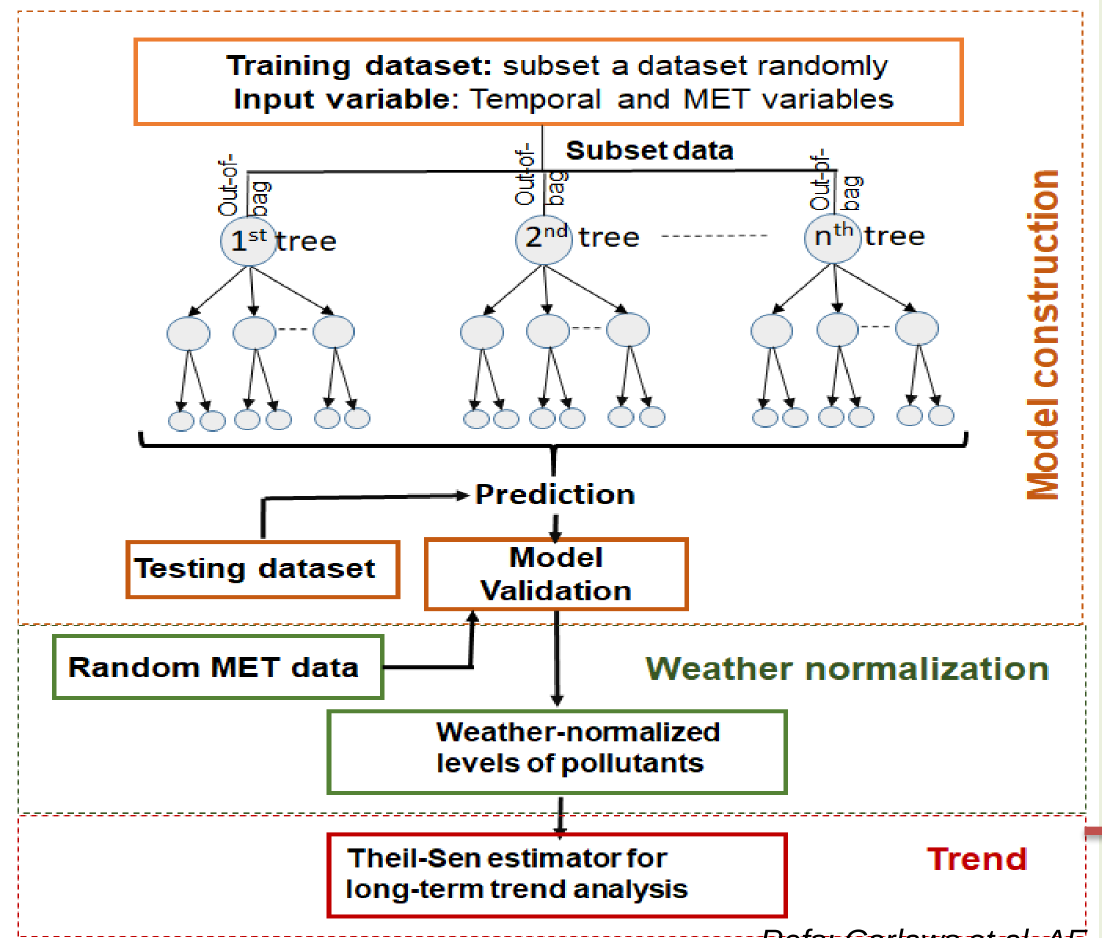
Long-term trend analysis method

The long-term time series of a pollutant can be split into components

$$\ln[C(t)] = C^{LT}(t) + C^{S}(t) + C^{STM}(t) + C^{WH}(t) + C^{WN}(t)$$

where, LT: long-term component; S: seasonal components; STM: Short-term component; WH: weekend/holiday impact; WN: white-noise is the residual.

A decision tree-based random forest technique



Training dataset: subset a dataset randomly
Input variable: Temporal and MET variables

Subset data

Prediction

Testing dataset

Model Validation

Model construction

Random MET data

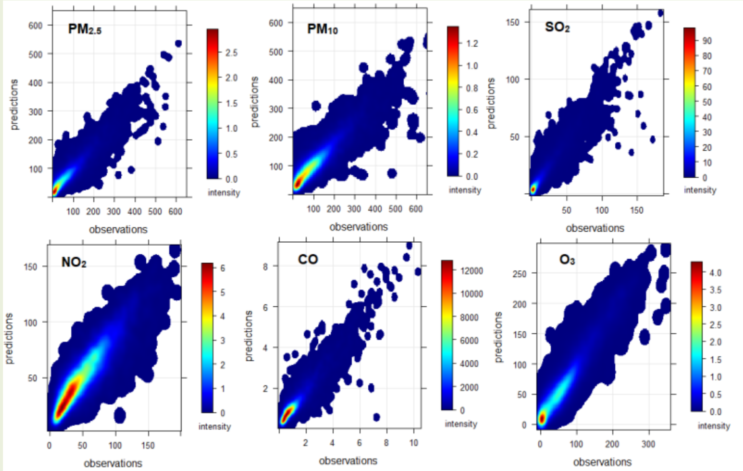
Weather normalization

Weather-normalized levels of pollutants

Theil-Sen estimator for long-term trend analysis

Trend

Model validation

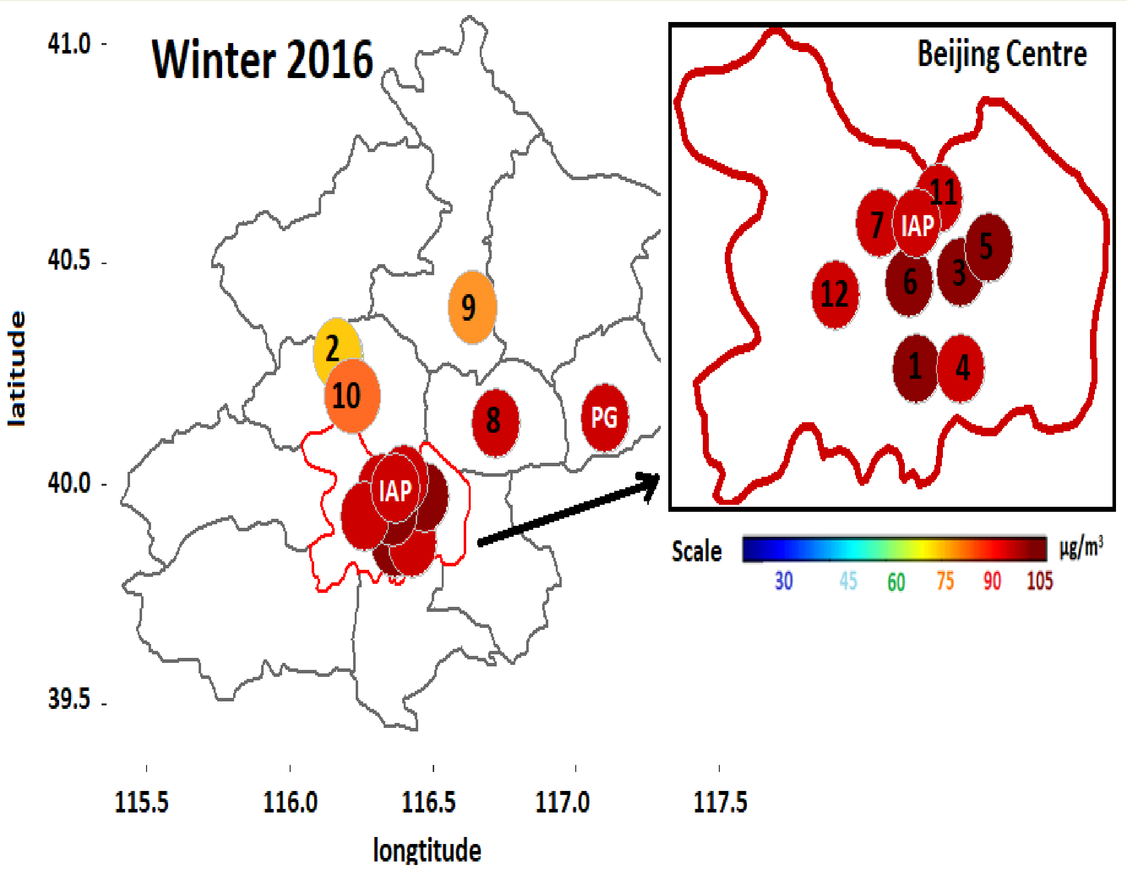


1. **List of policies**
2. **MEIC-emission inventory**

Refs: Carlaws et al. AE. 2009, Grange et al. ACP 2018, Vu et al . ACPD.2019

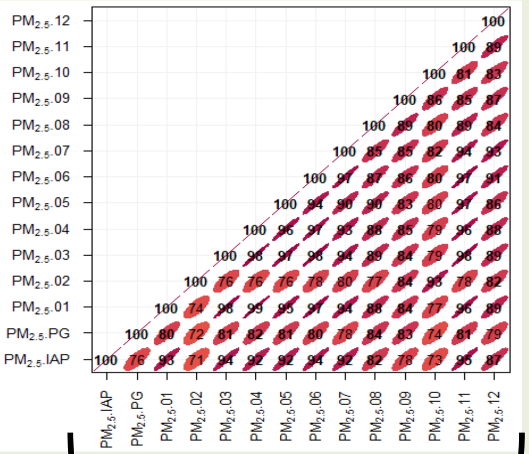
Input datasets of Air Pollutants in Beijing

Six key pollutants: **$PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3**
from 12 national monitoring stations during 2013-2017
& **30-year MET** data sets



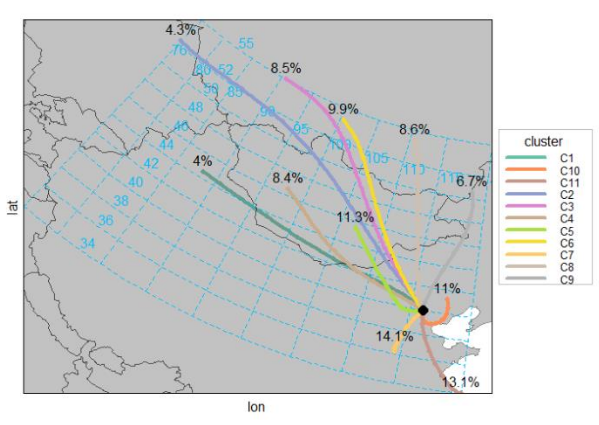
Spatial variation of $PM_{2.5}$ level during APHH winter campaign 2016

Refs: Shi et al 2019 ACP



	$PM_{2.5}$.IAP	$PM_{2.5}$.PG	$PM_{2.5}$.01	$PM_{2.5}$.02	$PM_{2.5}$.03	$PM_{2.5}$.04	$PM_{2.5}$.05	$PM_{2.5}$.06	$PM_{2.5}$.07	$PM_{2.5}$.08	$PM_{2.5}$.09	$PM_{2.5}$.10	$PM_{2.5}$.11	$PM_{2.5}$.12
$PM_{2.5}$.12														100
$PM_{2.5}$.11													100	89
$PM_{2.5}$.10												100	81	83
$PM_{2.5}$.09											100	86	85	87
$PM_{2.5}$.08										100	89	80	89	84
$PM_{2.5}$.07									100	85	85	82	94	93
$PM_{2.5}$.06								100	97	87	86	80	97	91
$PM_{2.5}$.05							100	94	90	90	83	80	97	86
$PM_{2.5}$.04						100	96	97	93	88	85	79	96	88
$PM_{2.5}$.03					100	98	97	98	94	89	84	79	98	89
$PM_{2.5}$.02				100	76	76	76	78	80	77	84	93	78	82
$PM_{2.5}$.01			100	74	98	99	95	97	94	88	84	77	96	89
$PM_{2.5}$.PG		100	80	72	81	82	81	80	78	84	83	74	81	79
$PM_{2.5}$.IAP	100	76	93	71	94	92	92	94	92	82	78	73	95	87

Urban/Suburban/Rural



30-year MET data sets
& back trajectories

Home key messages

- Understand the data first
- Practise basic coding more as you can
- How to use the techniques

Thank you for your attention