Queen Mary University of London

Instituto Politécnico Nacional

SAPIENS

SMART AIR POLLUTION INFORMATION ENABLING NEW SOLUTIONS

Analysis and Results

Marcella Bona (QMUL), **Nathan Heatley** (QMUL), Jia-Chen Hu (QMUL), Xiwen Zhang (QMUL), Adriana Lara (IPN), Alberto Luviano Juárez (IPN), John Moriarty (QMUL), Xiwen Zhang (QMUL)

IPN Graduate students: Carlos Jiménez González, Fernando Moreno-Gómez, Royce Richmond Ramírez Morales, Natan Ismael Vilchis-Tavera

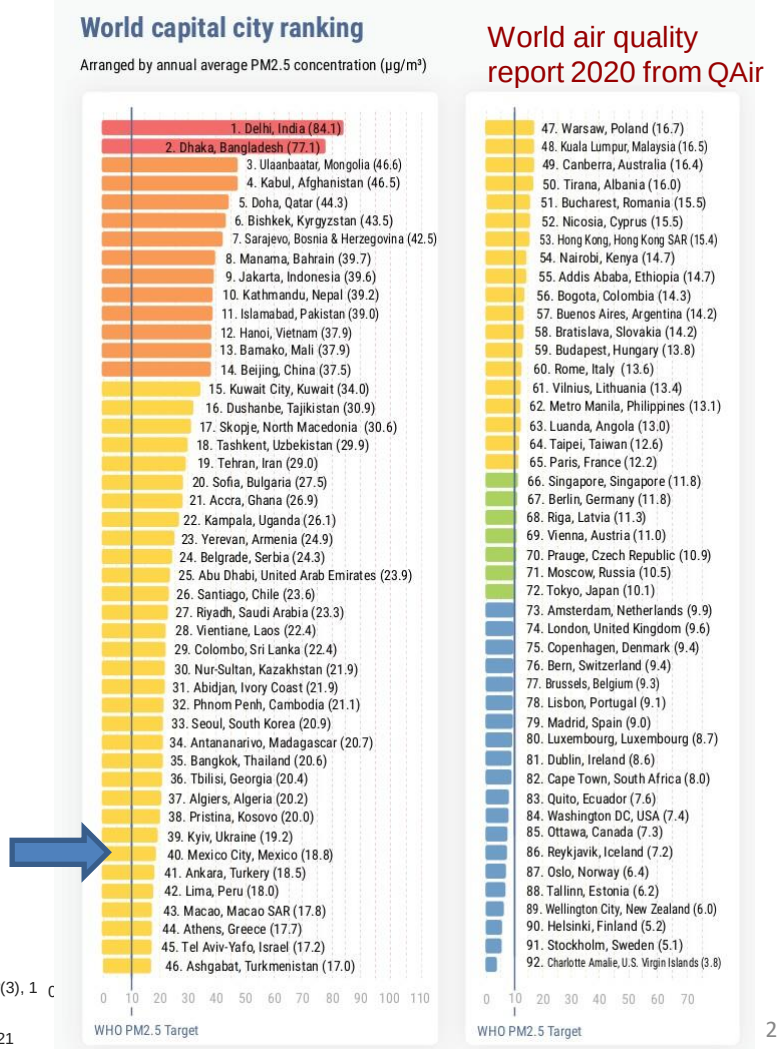
Nuclear Technologies talk

Motivations

Pollution has a devastating effect in our lives for those living in big cities.

The threshold for triggering alerts in Mexico City change every few years

- more relaxed than the current acceptable levels set by the WHO
- after 264 days of 2021, the city had reported just the 65% of days as "clean"



1. White, P. A., Gelfand, A. E., Rodrigues, E. R., & Tzintzun, G. (2019). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(3), 1 (1060.

2. Camahi, Elias, El aire de la ciudad de méxico supera con creces los límites que la OMS considera peligrosos para la salud, El País, 09/22/2021

3. www.iqair.com/Fworld-most-polluted-cities%2Fworld-air-quality-report-2020-en.pdf

Motivations

SAPIENS
SMART AIR POLLUTION INFORMATION ENABLING NEW SOLUTIONS

"Smart cities" can gather an enormous amount of data to help them improve the situation

Big data capabilities being built in many contexts allow to store, clean, and analyse these data.

Data can be analysed and with statistical modelling and machine learning we can learn behaviours and obtain predictions

"Smart cities" can imagine solutions and empower society to act, citizens and governments

Most data sets can come directly from the citizens themselves: traffic for example

Basic data: traffic data from google is free to download (up to some level)

SAPIENS: can we use the basic free level of traffic data to learn about pollution?

Data

Pollution data: CDMX Data Agency provides pollution data in terms of:

- 27 stations, levels of 9 pollutants recorded every hour
- Data Agency provides a clean set of data after 3 months of being taken

Traffic data: basic google images which are free and available instantaneously

- Downloaded in the SAPIENS database 3 times per hour.
- Traffic information google images of the 10 km^2 map around each of the sensors are downloaded.

Data processing

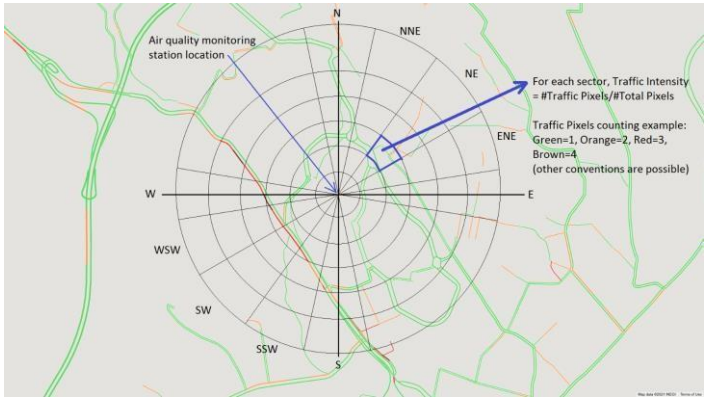
Pollution data: need an extra clean-up to remove outliers and to check the number of effective measurements for each sensor/station:

- Lots of sensors have very few data points
- Identified 3 sensors with high number of measurements for each pollutant
- The data from these 3 sensors are used in our modelling analysis
- The other sensors with a minimum of measurements to be used for validation

id_station_id	Null all Day	Null of 6 to 20 hrs
MER	489	170
CAM	962	485
PED	1541	779
IMP	2136	1246
TLA	2586	1391
ARA	4272	2492
LVI	4272	2492
VAL	4272	2492
SAG	7901	4607
SFE	10246	5990

Traffic data: images are translated into traffic intensity measurements based on concentric circles around the sensor position

- Information in terms of thickness of traffic colour-labelled lines or line segments: e.g. a wider street labelled orange is likely to have a different contribution than a narrower street labeled by the same traffic colour
- Count pixels of traffic colors to quantify traffic flow or volume.
- take other non-traffic pixels into consideration: if a station is located far from streets, the traffic intensity surrounding should be lower as those non-traffic pixels are not producing or “emitting” pollutants.



Pollutant Data Analysis

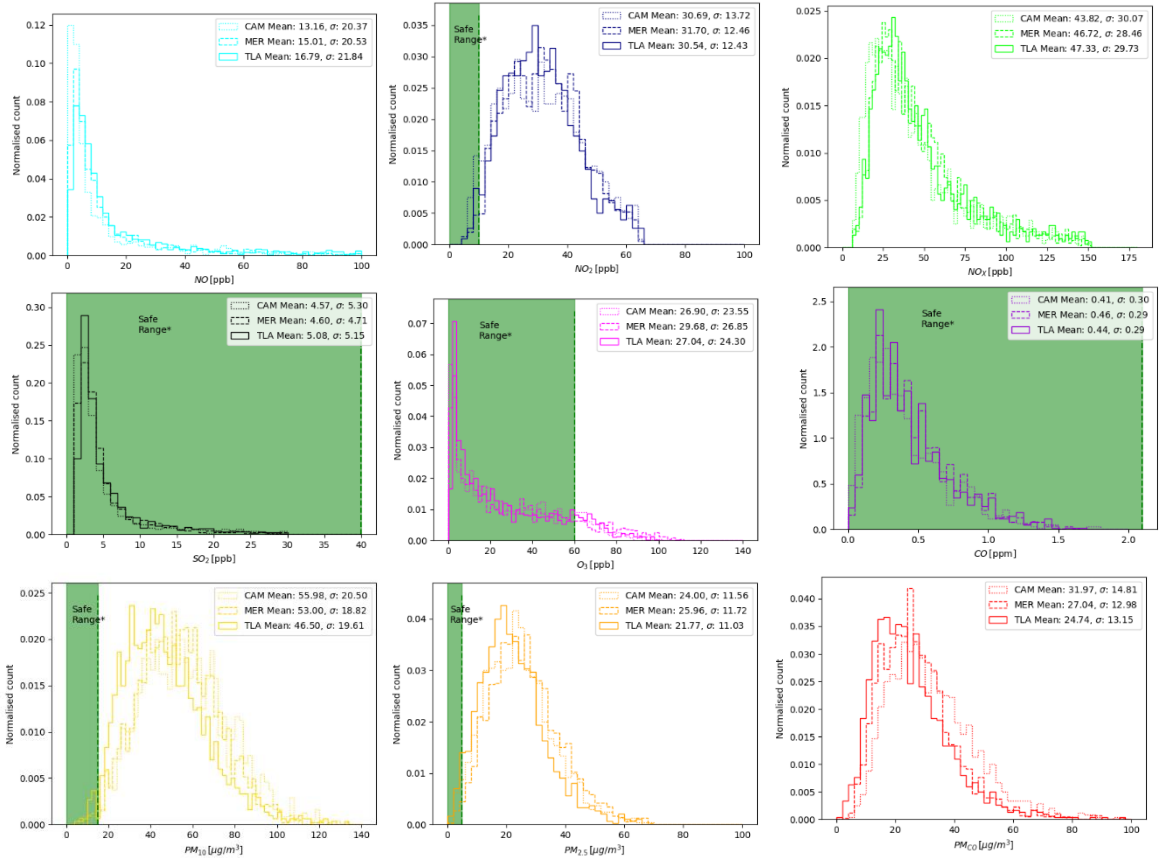
SAPIENS
SMART AIR POLLUTION INFORMATION ENABLING NEW SOLUTIONS

List of pollutants:

- PM10: particulate matter with 10 µm or less in aerodynamic diameter (µg/m³)
- PM2.5: particulate matter with 2.5 µm or less in aerodynamic diameter (µg/m³)
- PMCO: particulate matter with aerodynamic diameters between 2.5 and 10 µm (µg/m³)
- SO2: Sulfur Dioxide (ppb)
- O3: Ozone (ppb)
- CO: Carbon Monoxide (*ppm*)
- NO2: Nitrogen Dioxide (ppb)
- NO: Nitrogen Monoxide (ppb)
- NOX: Nitrogen Oxides (ppb)

µg/m³ = micrograms per cubic metre
ppb = parts per billion
ppm = parts per million

Pollutant Data Analysis



Distributions of the pollutant measurements:

Comparison between the three sensor stations

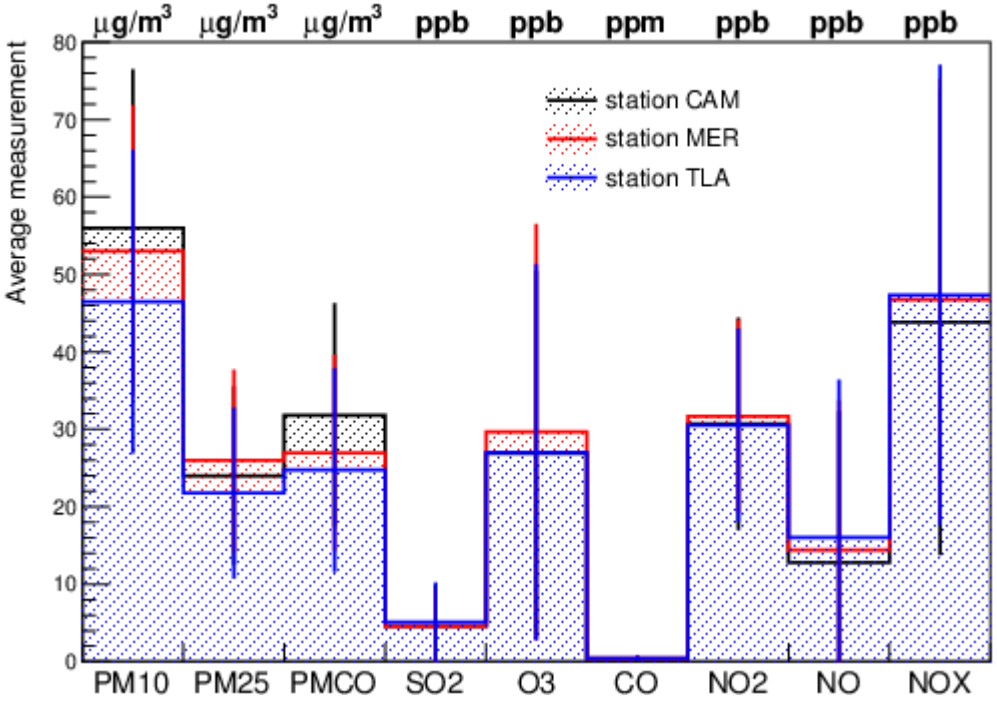
(labelled: CAM, MER and TLA)

Compatible distributions across the stations

Safe ranges for annual exposure shown in green, from the WHO [1]

1. https://www.who.int/publications/i/item/9789240034228

Pollutant Data Analysis



Average measurements for each pollutant and for each station used.

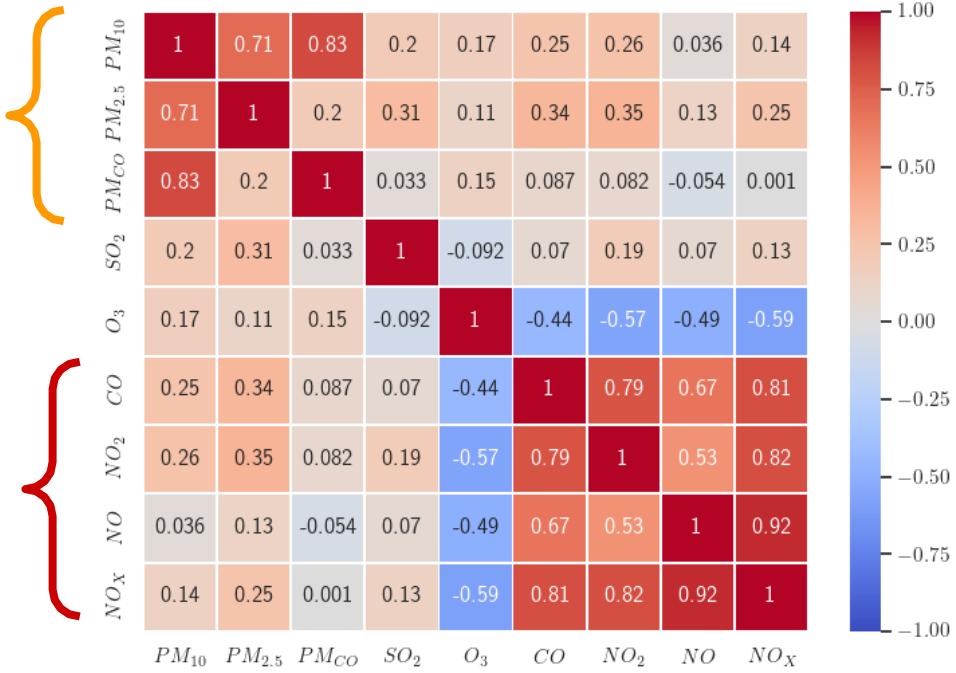
Error bars are the RMS of the pollutant distributions.

Units differ depending on the pollutant considered.

µg/m³ = micrograms per cubic metre; ppb = parts per billion; ppm = parts per million

Pollutant Data Analysis

Linear correlation coefficients



Correlation between pollutants:

Red colours: positive corr.
Blue colours: negative corr.

Two groups:
1: {PM10, PM25, PMCO}
2: {CO, NO2, NO, NOX}

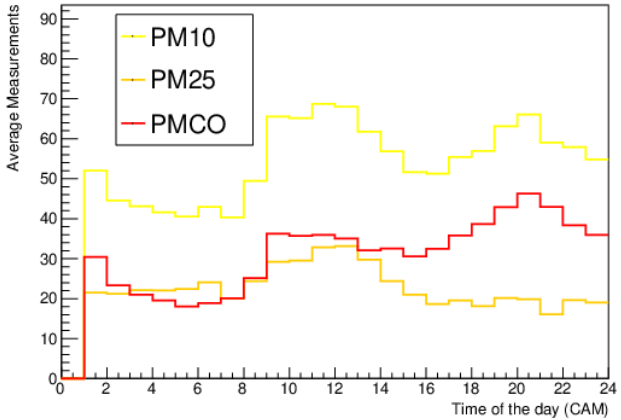
O3 anticorrelated with group 2

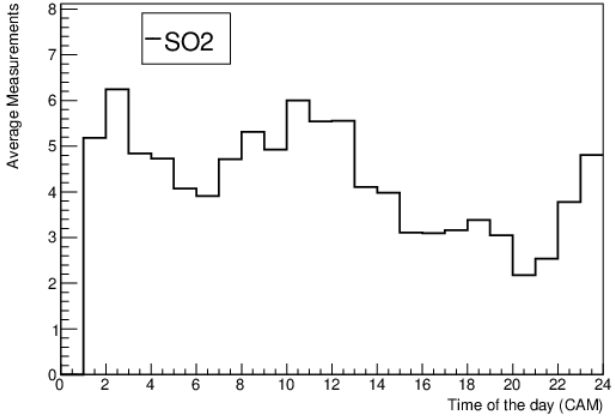
SO2 uncorrelated with the others

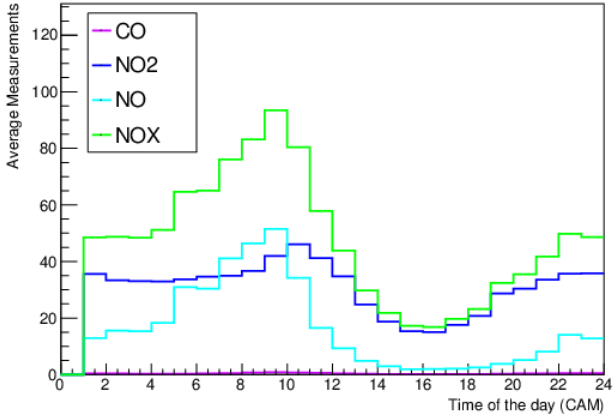
Pollutant Data Analysis

SAPIENS
SMART AIR POLLUTION INFORMATION ENABLING NEW SOLUTIONS

Check modulations during the day in the pollution measurements





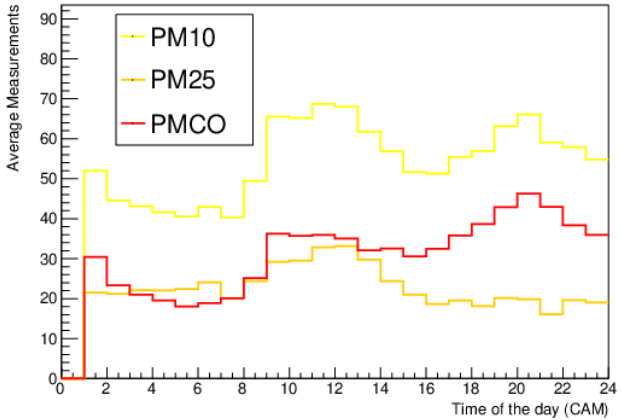


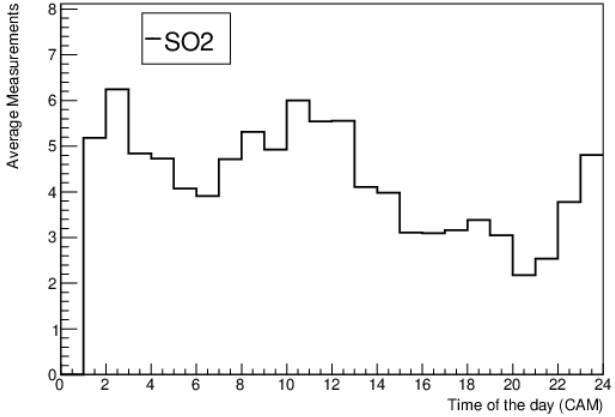
Pollutants in correlated groups with their absolute measurements.
Average measurements shown for each hour of the day.

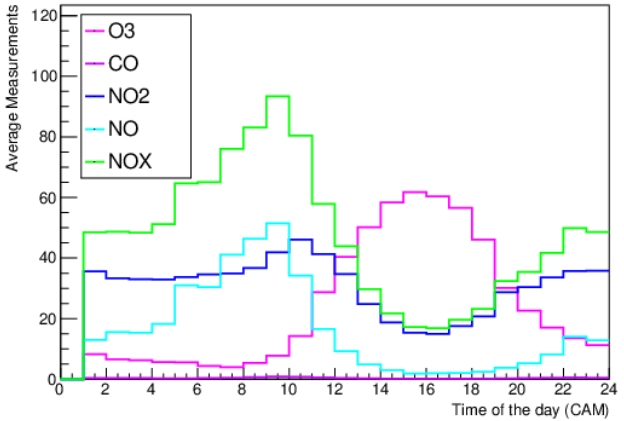
Pollutant Data Analysis

SAPIENS
SMART AIR POLLUTION INFORMATION ENABLING NEW SOLUTIONS

Check modulations during the day in the pollution measurements





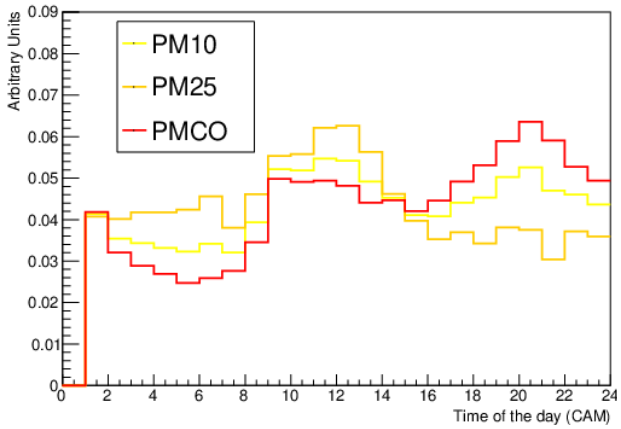


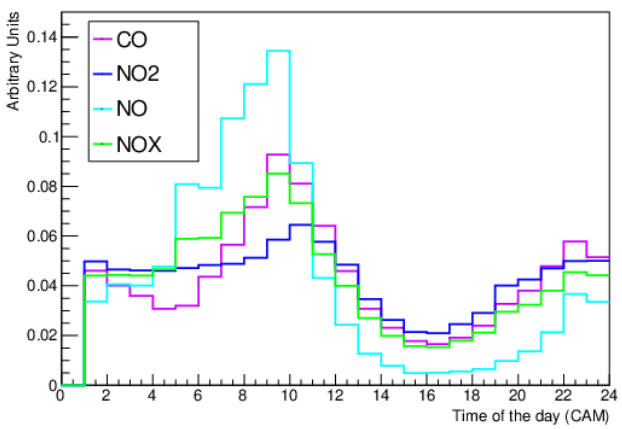
Adding anti-correlated O3

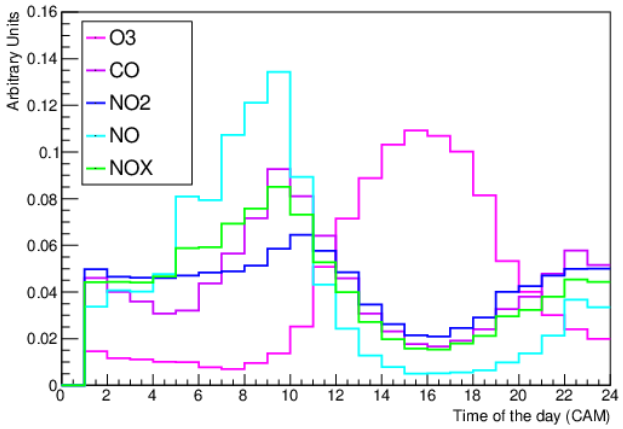
Pollutants in correlated groups with their absolute measurements.
Average measurements shown for each hour of the day.

Pollutant Data Analysis

Check modulations during the day in the pollution measurements







Pollutants in correlated groups:
here normalised distributions for shape comparison.

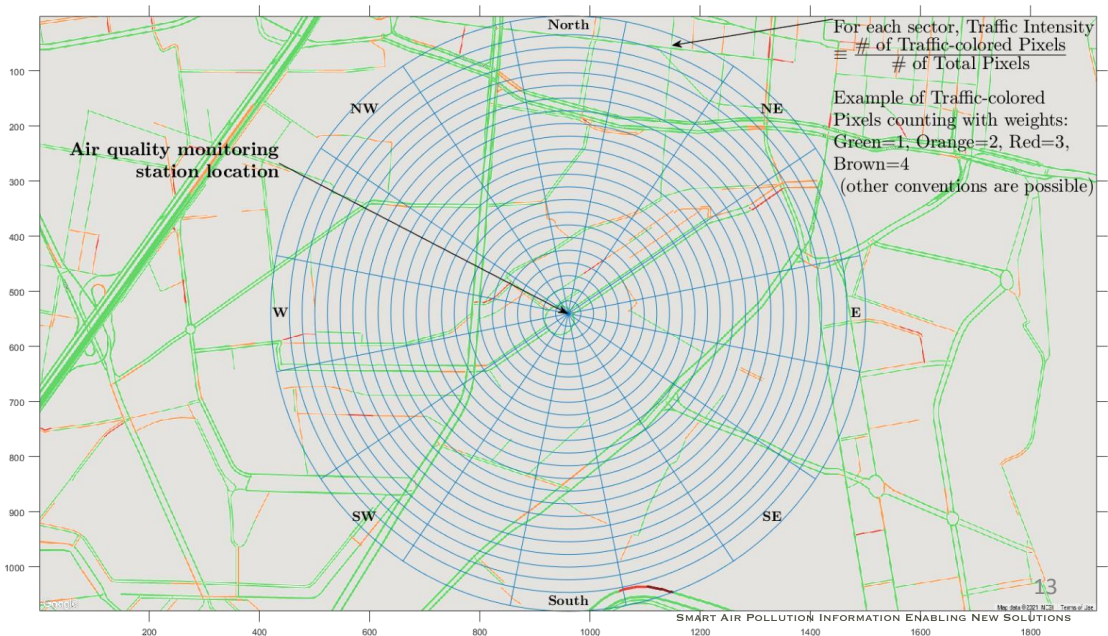
Traffic Intensity Model

Google Maps images with traffic layer/colouring: green, orange, red, dark red
% of coloured pixels in annular sectors -> traffic intensity; physical density “field”

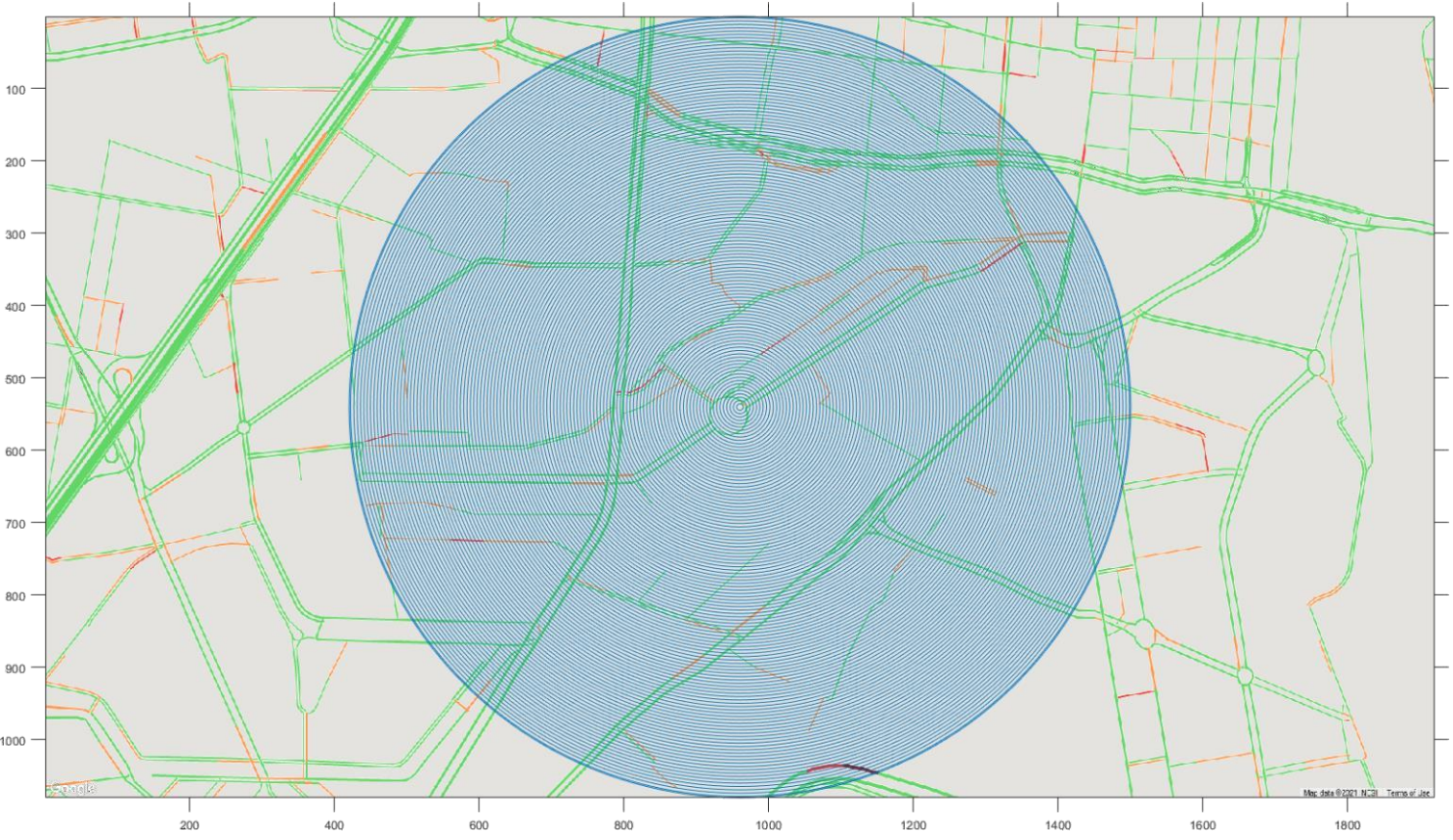
Dimension reduction
from HD image 1920x1080:

(4 colours)x(16 angles)x(118 rings)
Or aggregating the angles:
(4 colours)x(118 rings)

Only 23 rings in this plot
118 rings shown in the next page



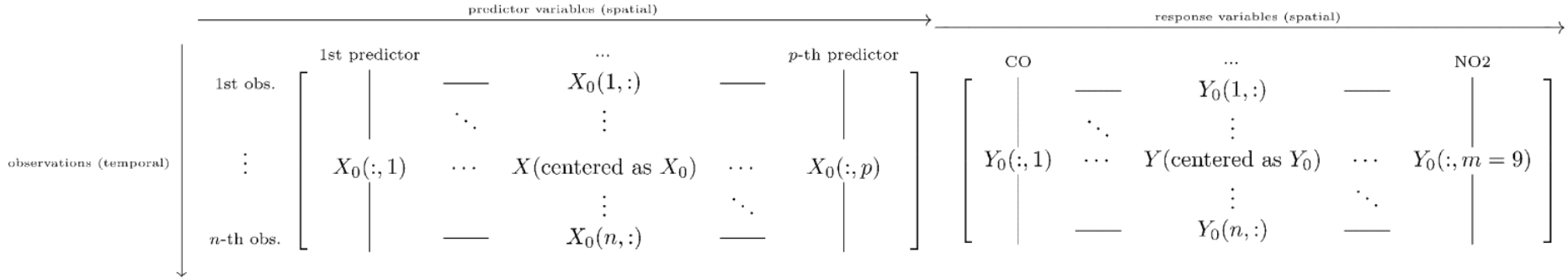
Traffic Intensity Model



Width of 118 concentric rings: 10m

Regression Modeling: data matrices

Spatio-Temporal Data Matrices X (traffic predictors; centered as X_0 with 0 mean & rescaled with Var=1) and Y (pollutants responses; centered & rescaled as Y_0):



where p = (4 colors)×(118 rings) (or (4 colors)×(16 angles)×(118 rings)) traffic predictor variables, $m = 9$ pollutant response variables;

n observations, depending on station/sensor (each has different # of missing/null readings)

Regression Modeling: PLSR

Purpose of our modeling is two-fold:

(1) (interpretations) get traffic activities $\xrightarrow{\text{mapping}}$ pollutant concentrations, and reveal insight on their detailed relations, and

(2) (predictions) predict pollutant concentrations based on traffic activities.

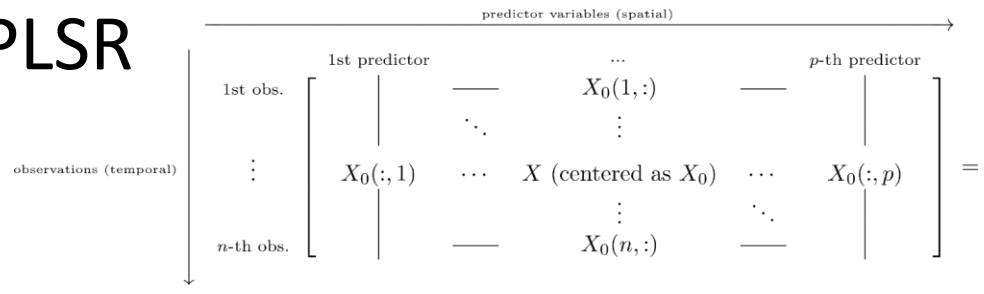
Most black-box machine learning techniques are good at (2) only;

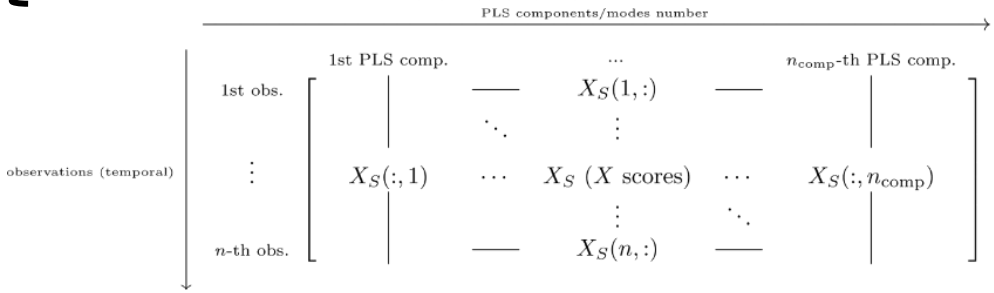
Partial least squares (PLS) regression formulation:

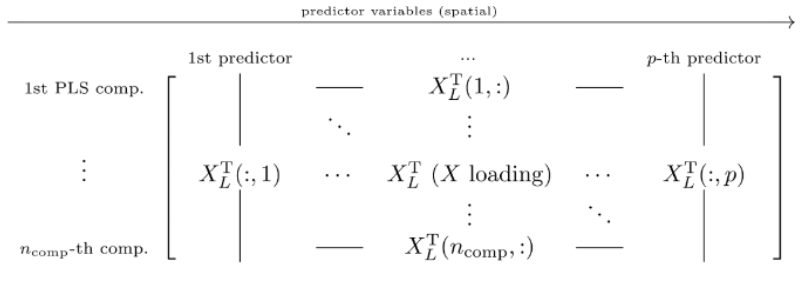
$$\begin{cases} X_0 = X_S X_L^{\mathsf{T}} + X_{\text{residuals}} \\ Y_0 = Y_S Y_L^{\mathsf{T}} + Y_{\text{residuals}} \end{cases}$$

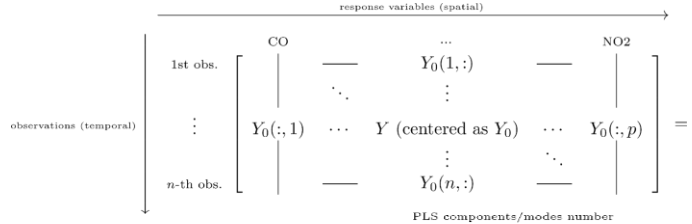
Regression Modeling: PLSR

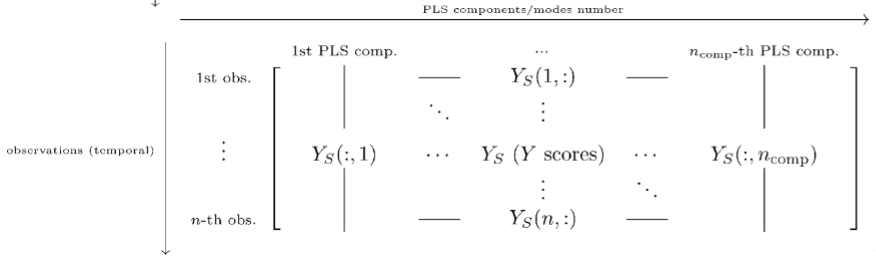
$$\begin{cases} X_0 = X_S X_L^{\mathsf{T}} + X_{\text{residuals}} \\ Y_0 = Y_S Y_L^{\mathsf{T}} + Y_{\text{residuals}} \end{cases}$$

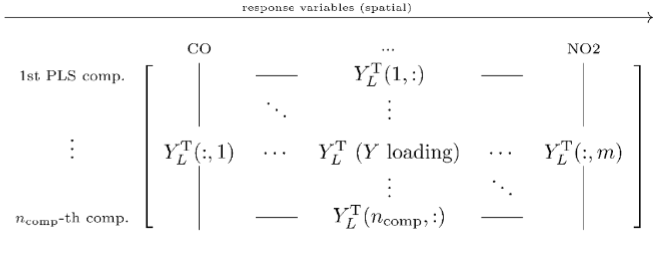




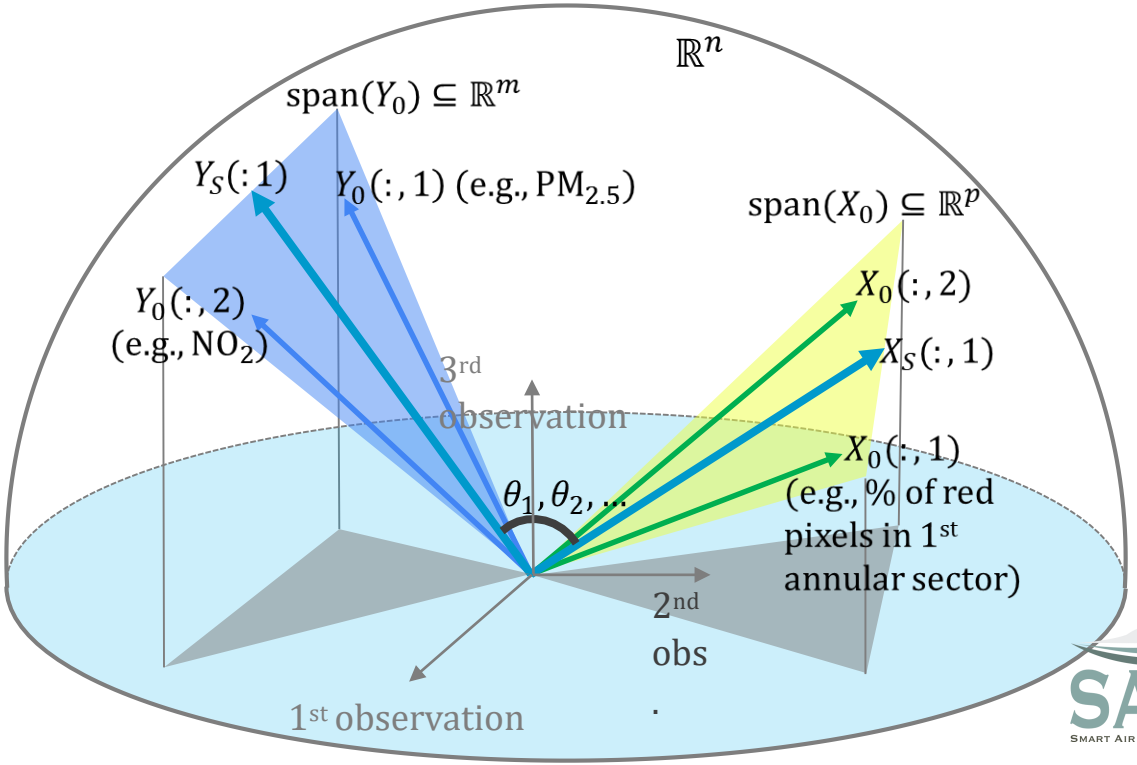








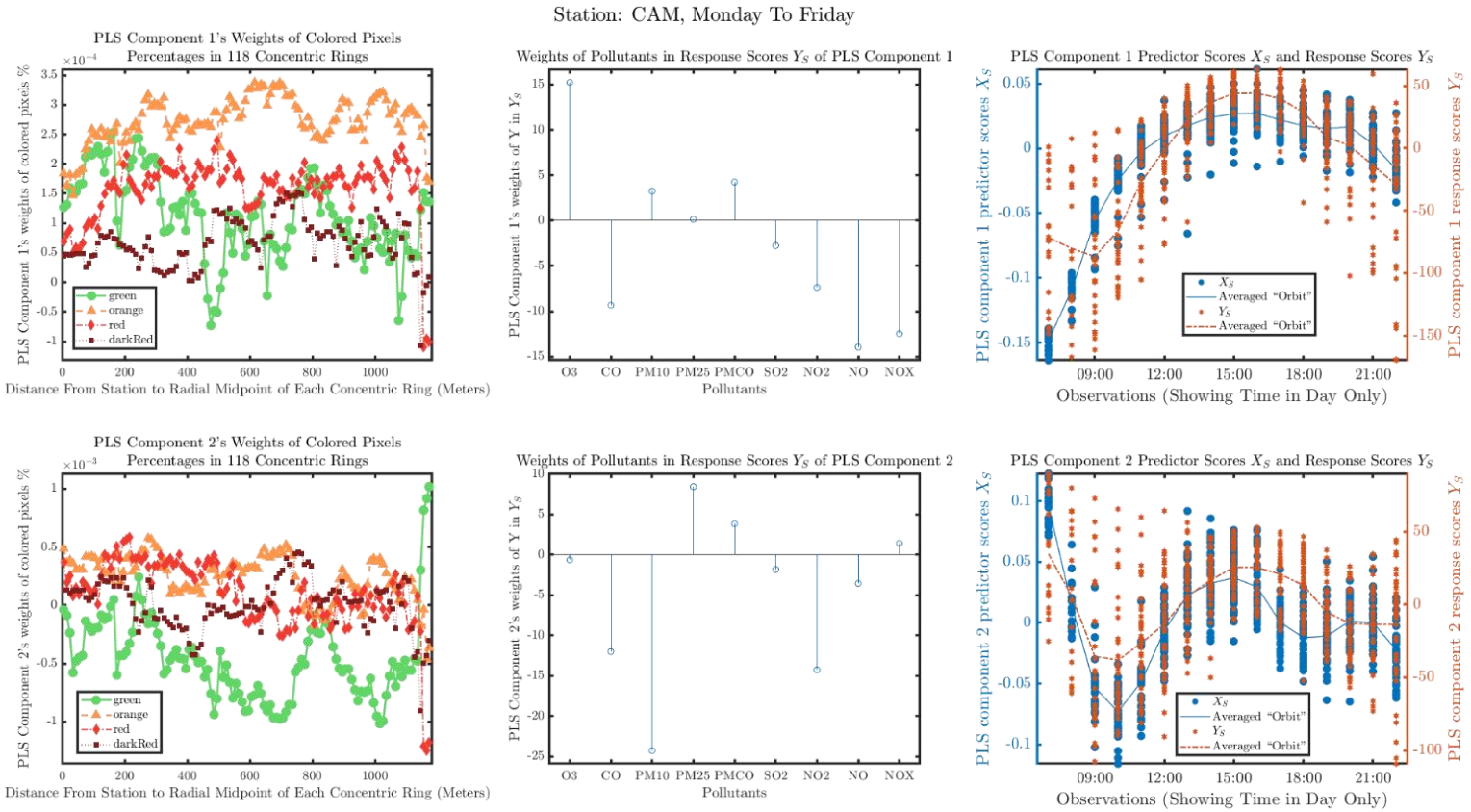
PLS Regression and geometric interpretation



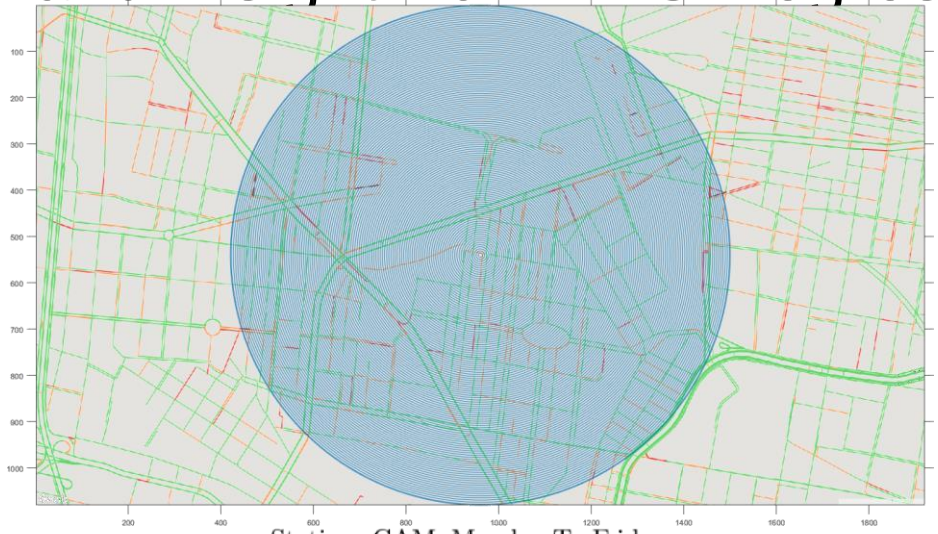
SAPIENS
SMART AIR POLLUTION INFORMATION ENABLING NEW SOLUTIONS

Interpretations and Insight from PLS Regression Modeling

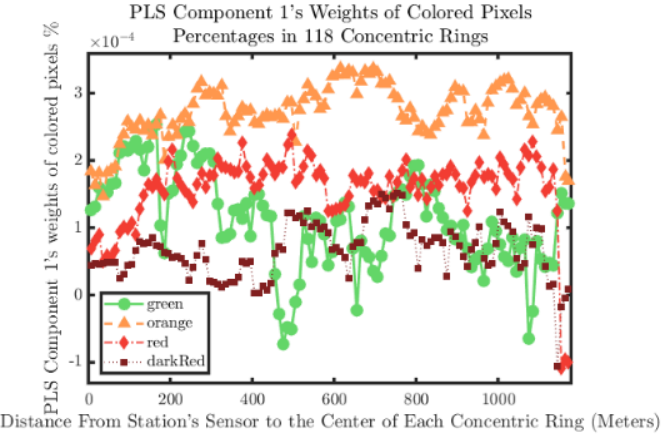
1st PLS component has physically meaningful interpretation:

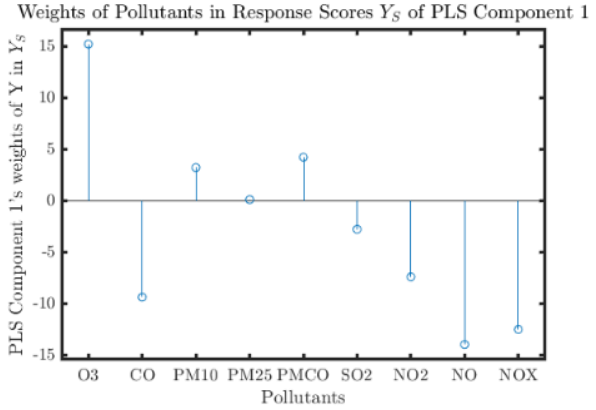


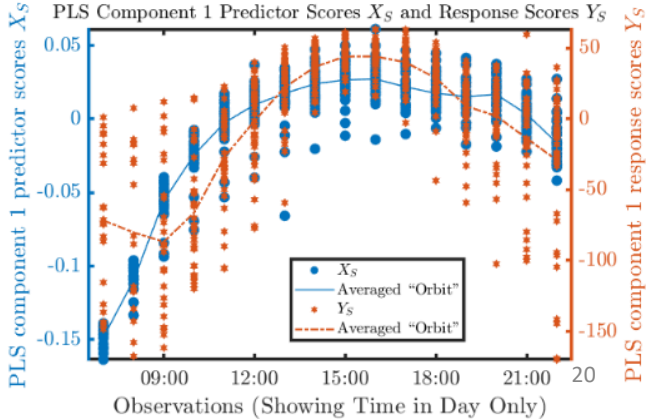
Interpretations and Insight from PLS Regression Modeling



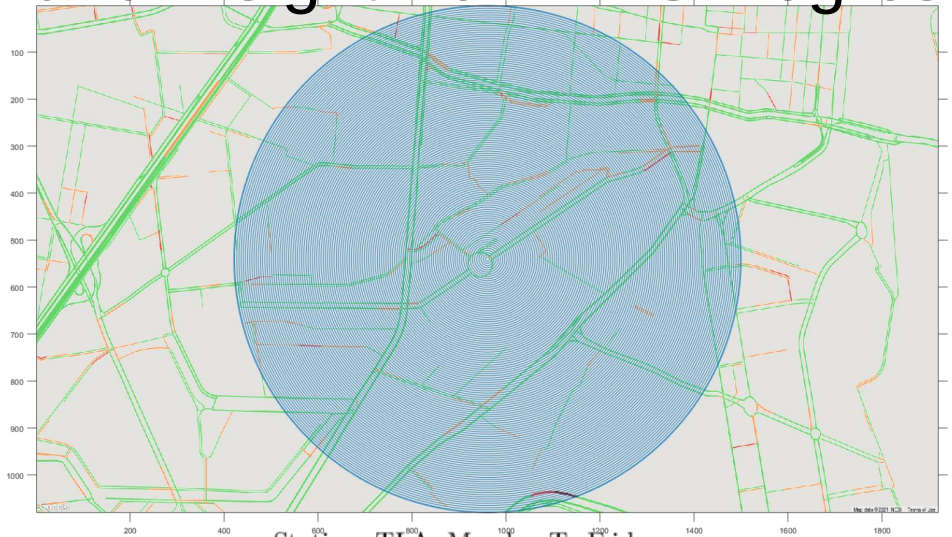
Station: CAM, Monday To Friday



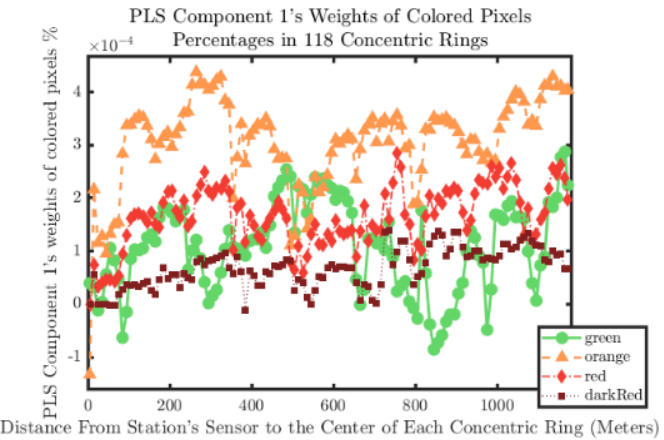


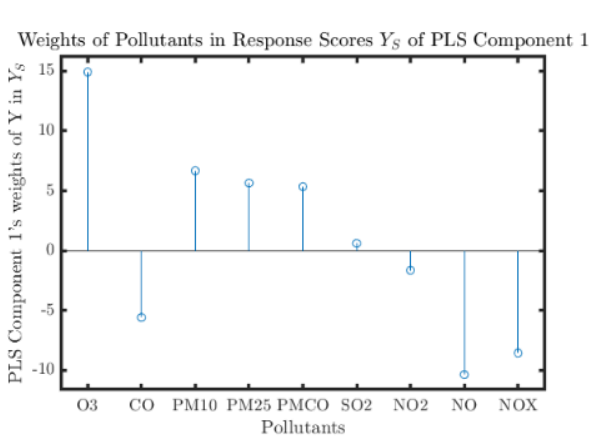


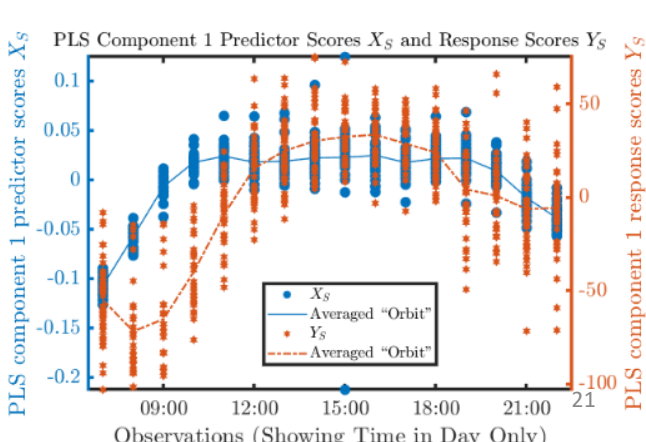
Interpretations and Insight from PLS Regression Modeling



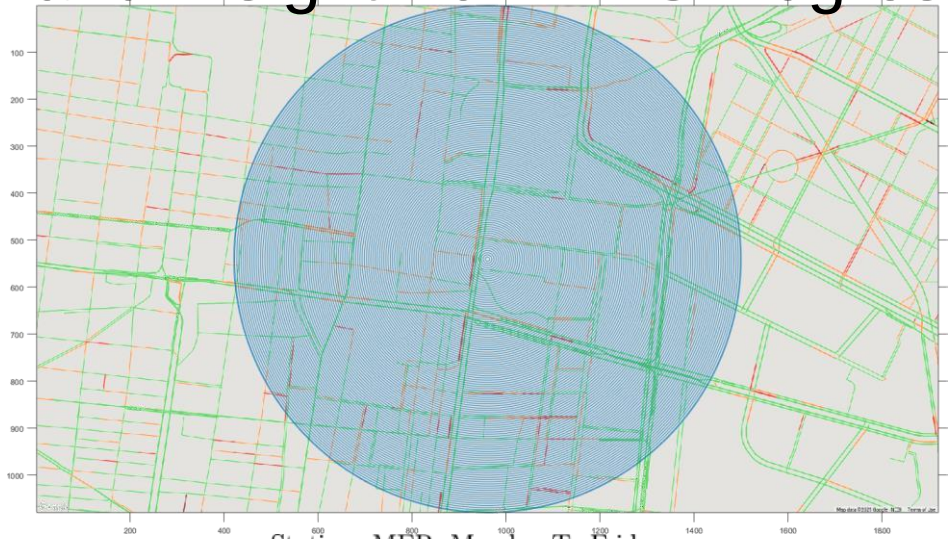
Station: TLA, Monday To Friday



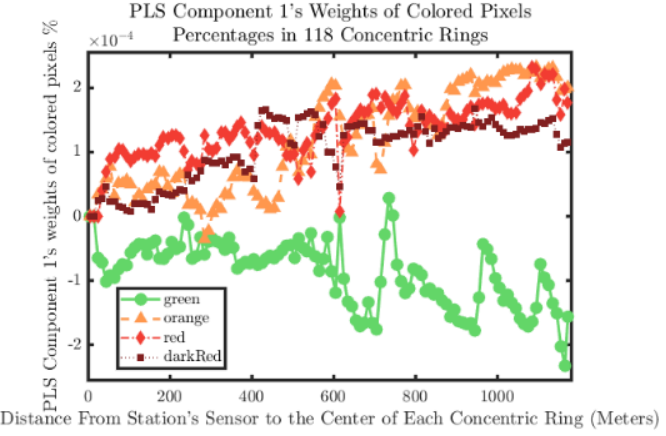


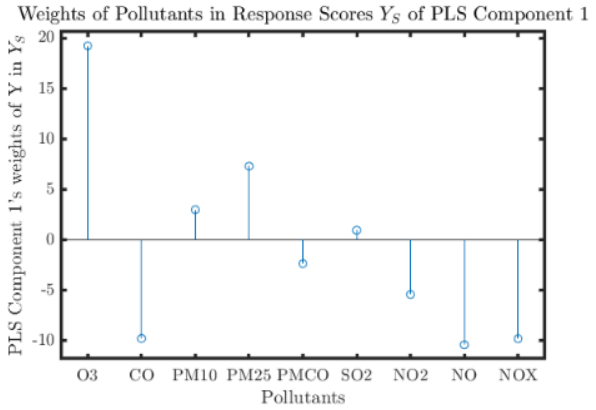


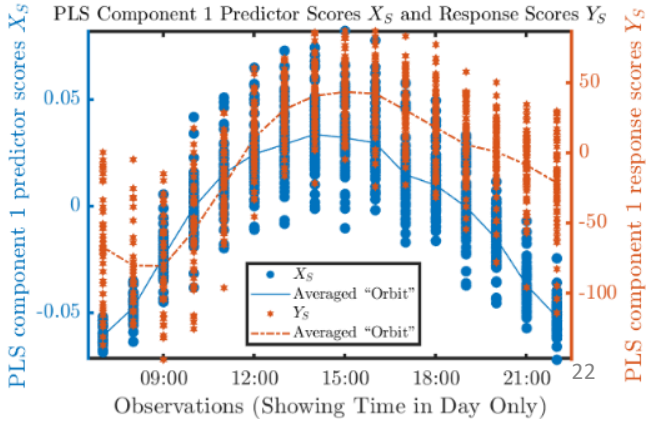
Interpretations and Insight from PLS Regression Modeling



Station: MER, Monday To Friday







PLSR Modeling and Prediction Performance

In actual model fitting, a (p predictors)-by-(m responses) coefficients matrix $\beta_{n_{\text{comp}}}$ is fitted in least squares sense for

$$Y_S Y_L^{\mathsf{T}} = X_S X_L^{\mathsf{T}} \beta_{n_{\text{comp}}}$$

using truncated n_{comp} PLS components, and $X_S X_L^{\mathsf{T}}$ and $Y_S Y_L^{\mathsf{T}}$ are "reconstruction" of X_0 and Y_0

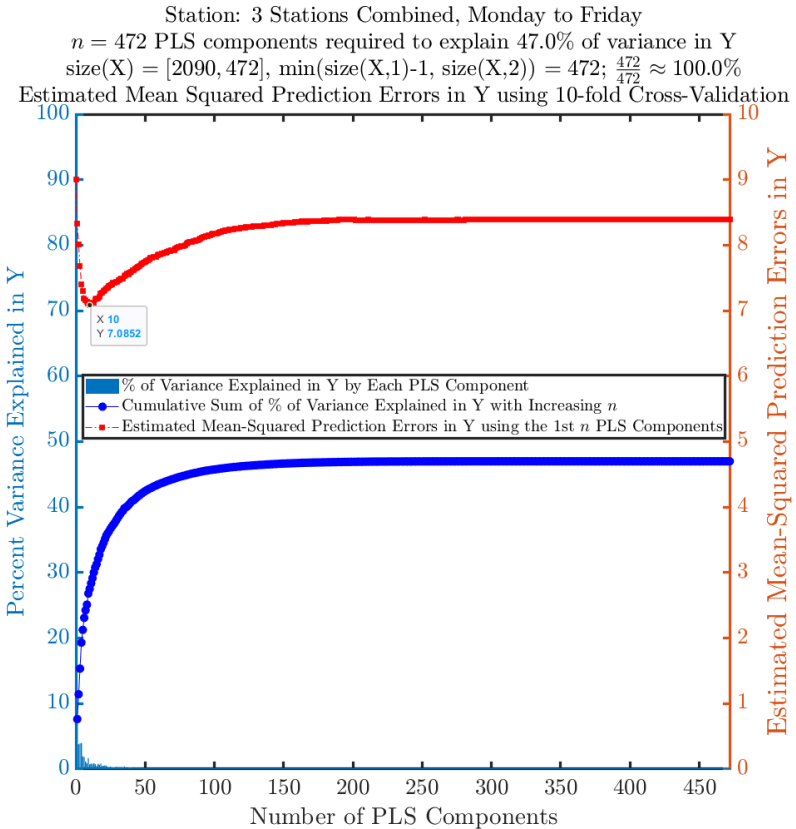
Spatial dimension in X_0 reduced from p to n_{comp}, effectively fitting $Y_S = X_S \beta$ (a "partial" least-square)

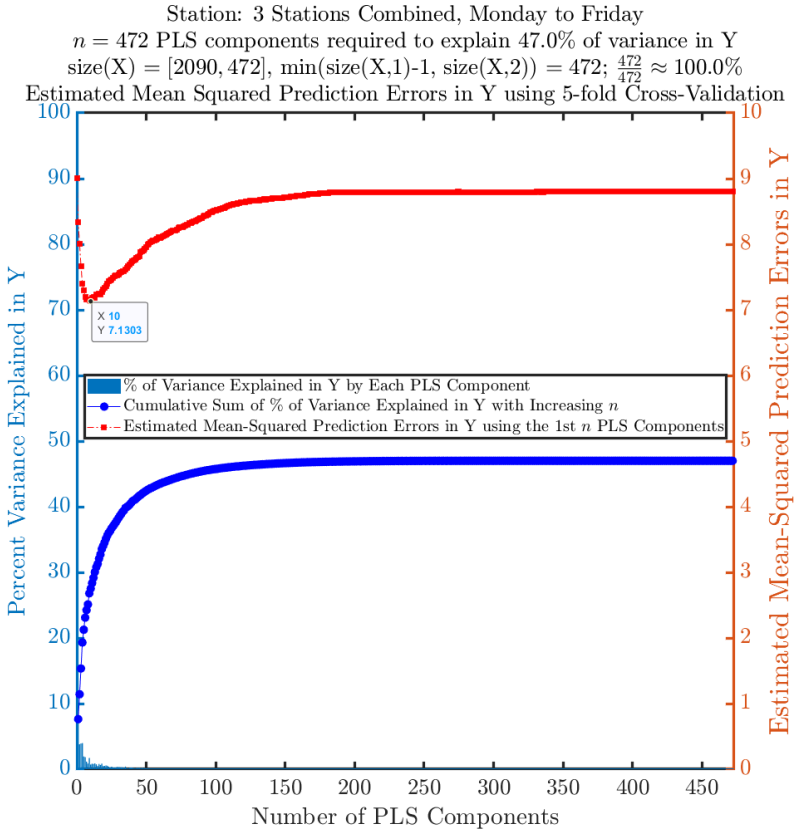
n_{comp} can be fixed by cross-validation to minimize the expected mean-squared errors (MSE) $\left(Y_0 - Y_S Y_L^{\mathsf{T}}\right)^2$

(a preliminary result on next page)

PLSR Modeling and Prediction Performance

(all 9 pollutant response variables are centered to 0 mean with rescaled Var=1)





Outlook and Conclusions

SAPIENS has built a database with both pollution measurements and traffic images, so we have:

- Cleaned and analysed the data and identified patterns
- Developed a model to extract the traffic intensities from Google Map images
- Used the regression modeling to (1) obtain interpretable insights on the relation between traffic and pollutants; and (2) train it on the data from three stations (traffic and pollution data) and cross-validated it to avoid overfitting

On-going activities:

- Validation/testing phase: use other sensors data to validate/test model
- Paper in preparation

Outlook and Conclusions

There are more ideas and more possibilities to exploit and learn from these data.

More ideas on how to exploit the predicting power of the modeling

E.g., incorporating meteorological data, going beyond linear modeling techniques, etc.