Data analysis on community $PM_{2.5}$ hot-spot identification and quantification (R code)

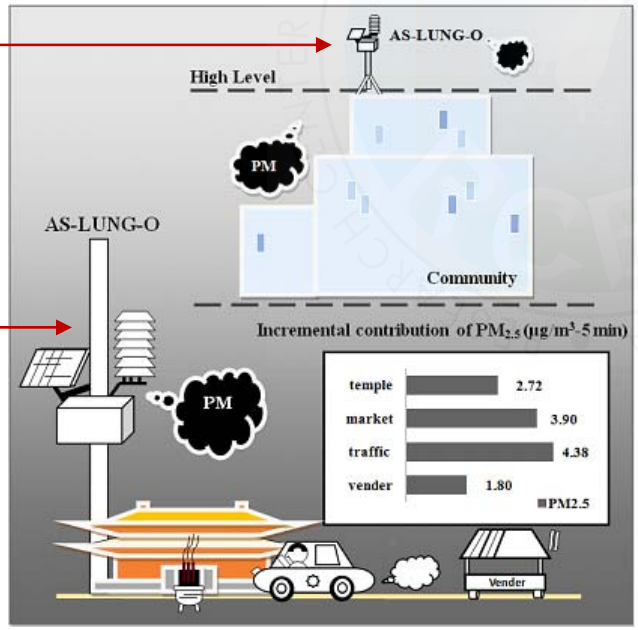
SC Candice Lung & WC Vincent Wang

ADVANCED INSTITUTE ON Hi-ASAP (2020)
ACADEMIA SINICA, TAIWAN

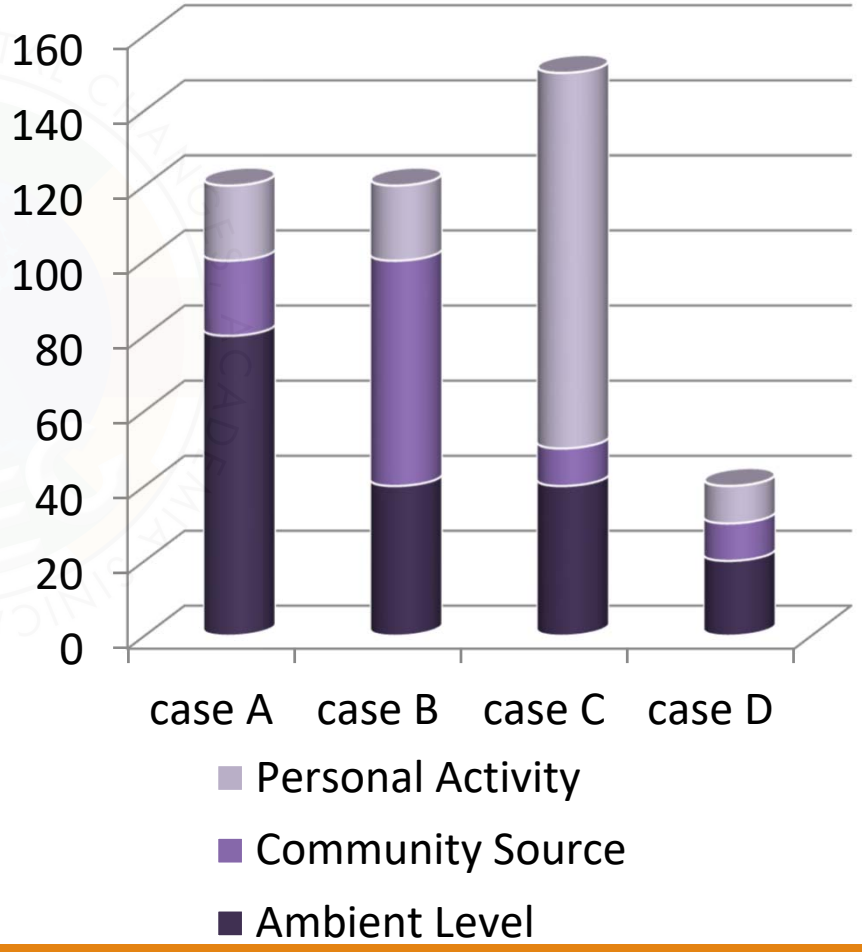
Exposure Concept

Exposure Level =**Ambient** level
+ **Community source** contribution
+ **Personal activity** contribution

Ambient level



Street level
AS-LUNG



A Versatile Low-cost Sensing Device for Assessing $PM_{2.5}$ Spatiotemporal Variation and Quantifying Source Contribution

Shih-Chun Candice Lung[a,b,c,*], Wen-Cheng Vincent Wang[a], Tzu-Yao Julia Wen[a], Chun Hu Liu[a], Shu-Chuan Hu

Science of the Total Environment, 716.

DOI: 10.1016/j.scitotenv.2020.137145.

This study aims:

(1) to evaluate the applicability of AS-LUNG-O to real-time $PM_{2.5}$ monitoring in a mountain community; and

(2) to quantify $PM_{2.5}$ contributions of common Asian community sources to human exposure or community air quality.

Science of the Total Environment 716 (2020) 137145

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

A versatile low-cost sensing device for assessing $PM_{2.5}$ spatiotemporal variation and quantifying source contribution

Shih-Chun Candice Lung [a,b,c,*], Wen-Cheng Vincent Wang [a], Tzu-Yao Julia Wen [a], Chun-Hu Liu [a], Shu-Chuan Hu [a]

[a] Research Center for Environmental Changes, Academia Sinica, Taipei, Taiwan
[b] Department of Atmospheric Sciences, National Taiwan University, Taipei, Taiwan
[c] Institute of Environmental Health, National Taiwan University, Taipei, Taiwan

HIGHLIGHTS

- AS-LUNG-O is a versatile low-cost sensing device, capable of conducting research.
- AS-LUNG-O can be operated for the street-level monitoring.
- Highest one-min $PM_{2.5}$ peak at a street site was 36.2 times of that at 10-m height.
- $PM_{2.5}$ increments from stop-and-go were higher than those of passing-by vehicles.
- Cooking and incense-burning are important $PM_{2.5}$ community sources.

GRAPHICAL ABSTRACT

ARTICLE INFO

Article history:
Received 22 November 2019
Received in revised form 18 January 2020
Accepted 4 February 2020
Available online 5 February 2020

Editor: Pavlos Kassomenos

Keywords:
PM micro-sensor
Community air quality
Low-cost sensor
Real-time $PM_{2.5}$ monitoring
Community source evaluation

ABSTRACT

This study evaluated a newly developed sensing device, AS-LUNG-O, against a research-grade GRIMM in laboratory and ambient conditions and used AS-LUNG-O to assess $PM_{2.5}$ spatiotemporal variations at street levels of an Asian mountain community, which represented residents' exposure (at the interface of atmosphere and human bodies leading to potential health impacts). In laboratory, R^2 of 1-min AS-LUNG-O and GRIMM was 0.95 ± 0.04 (n = 64,179 for 40 sets). After conversion with individual correction equations, their correlation in ambient tests was 0.93 ± 0.05, with absolute % difference of only 10 ± 9%. Ten AS-LUNG-O sets were installed at street sites with another one at 10 m above ground on July 1–28 and December 2–31, 2017 in Nantou, Taiwan. Important source contributions to $PM_{2.5}$ were quantified with regression analysis. Temporal variation expressed as the daily max/mean of 5-min $PM_{2.5}$ reached 13.7 in July and 12.2 in December. Spatial variation expressed as the percent coefficients of variance (%CV) across ten community locations was 22% ± 20% (max: 199%) in July and 19 ± 18% (max: 206%) in December. Incremental contribution from the stop-and-go traffic, market, temple, and fried-chicken vendor to $PM_{2.5}$ at 3–5 m away were 4.38, 3.90, 2.72, and 1.80 μg/m^3, respectively. Significant spatiotemporal variations and community source contributions revealed the importance of assessing neighborhood air quality for public health protection. For long-term air quality monitoring, the percentage of available power

* Corresponding author at: Research Center for Environmental Changes, Academia Sinica, No. 128, Sec. 2, Academia Rd, Nangang, Taipei 11529, Taiwan.
E-mail address: sclung@rcec.sinica.edu.tw (S.-C.C. Lung).

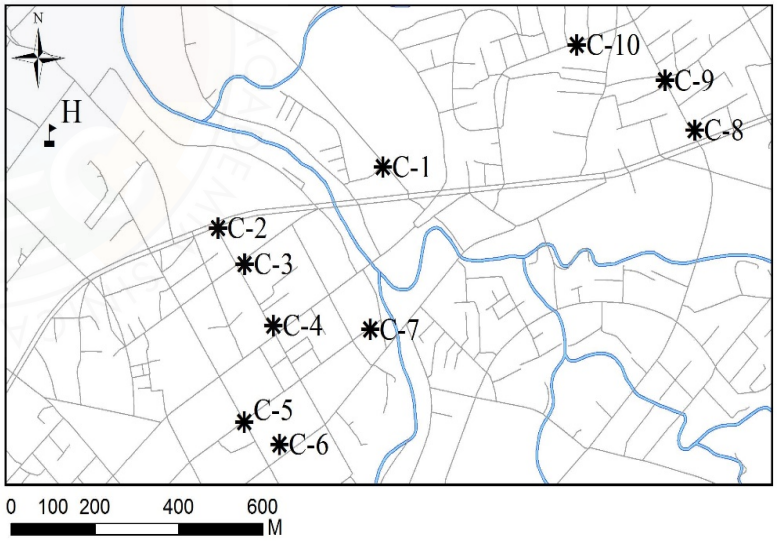
https://doi.org/10.1016/j.scitotenv.2020.137145
0048-9697/

Monitoring Strategy

10 AS-LUNG-O devices were placed at 2.5 meters above ground near certain community sources; one was for high-level site (10 m above ground). Two periods of measurements were conducted in July and December, 2017.

Site	Pollution sources
C-1	School, traffic type 1
C-2	Traffic type 1
C-3	Market, traffic type 1
C-4	Gas station, traffic type 1
C-5	Vendor, traffic type 1
C-6	Temple, traffic type 2
C-7	Street background
C-8	Traffic type 2
C-9	Temple, traffic type 2
C-10	Temple, traffic type 2

Traffic type 1: Traffic with passing-by vehicles
Traffic type 2: Stop-and-go traffic (stop near the traffic light)



Multiple Regression Equation

$$P_{site} = \sum \beta i Xi + \gamma 1(P_{high\text{-}level}) + \gamma 2\ (ws) + \gamma 3\ (temperature) + \gamma 4\ (rh) + \gamma 5\ season + \beta 0 + \varepsilon$$

P_{site}: 5-min $PM_{2.5}$ at the street-level site.

$P_{high\text{-}level}$ is 5-min $PM_{2.5}$ at the high-level site.

Xi: a dummy variable for pollution source. Xi is assigned as 1 if there is one of those pollution sources within 3-5 m of the site. If there is no pollution source, Xi is assigned as 0.

ws, temperature, and rh: ambient wind speed, temperature, and relative humidity.

season: a dummy variable for season. Summer (Jul.) is assigned as 0 and winter (Dec.) is assigned as 1.

$\gamma 1$, $\gamma 2$, $\gamma 3$, $\gamma 4$ and βi: regression coefficients.

$\beta 0$: the intercept.

ε: an error term.

Source Code in R (1/6)
Read the data file

The pound sign, #, is used for annotations or comments in R. You may write down some notes for your own reference. After this sign, the text will not be run.

The template data file is in the "input" folder.

Line	Script
1	#read data from the "input" folder
2	data_array <- read.csv(file='./input/data_community_workshop.csv')

The variable, data_array, is used for the storage of data, which are read from the data file in the directory indicated in the right side.

The arrow sign, <-, is used to assign data to the variable. Data is in the right side; the variable is in the left side.

The function, read.csv(), is used to read data from the 'csv' file. The parameter, 'file= ', is used to assign the path of the data file. The path of the data file, './input/data_community_workshop.csv', has to be put in middle of the quote signs.

Data Frame of the Template File

(example file, 1/6 of the data in the paper)

time: Data time,
site: The site where AS-LUNG-O was,
site_pm2.5: $PM_{2.5}$ measurements of street-level AS-LUNG-O,
high_level_pm2.5: $PM_{2.5}$ measurements of high-level AS-LUNG-O (ambient level),
ws: Ambient wind speed,
temperature: Ambient temperature,
rh: Ambient relative humidity.

	A	B	C	D	E	F	G
1	time	site	site_pm2.5	high_level_pm2.5	ws	temperature	rh
2	2017/7/6 06:00	C_1	22.3	17.3	0.0	25.0	92.8
3	2017/7/6 06:05	C_1	21.6	16.8	0.0	25.3	91.8
4	2017/7/6 06:10	C_1	22.0	16.2	0.0	25.7	91.1
5	2017/7/6 06:15	C_1	21.5	19.1	0.0	25.8	90.2
6	2017/7/6 06:20	C_1	20.9	19.6	0.0	25.6	91.8
7	2017/7/6 06:25	C_1	20.4	19.0	0.0	25.8	91.7
8	2017/7/6 06:30	C_1	18.8	17.8	0.0	25.9	91.0
9	2017/7/6 06:35	C_1	19.9	17.4	0.0	26.4	90.5
10	2017/7/6 06:40	C_1	20.5	16.5	0.0	27.2	88.5
11	2017/7/6 06:45	C_1	21.7	15.6	0.0	26.9	86.8
12	2017/7/6 06:50	C_1	19.1	16.7	0.0	26.8	87.1
13	2017/7/6 06:55	C_1	17.7	16.4	0.0	26.9	87.6
14	2017/7/6 07:00	C_1	16.1	14.8	0.0	27.6	85.6
15	2017/7/6 07:05	C_1	15.6	14.0	0.0	27.6	84.7
16	2017/7/6 07:10	C_1	15.5	11.9	0.1	27.6	84.1
17	2017/7/6 07:15	C_1	14.6	12.2	0.0	28.4	83.2
18	2017/7/6 07:20	C_1	15.9	11.9	0.0	28.7	81.4
19	2017/7/6 07:25	C_1	12.6	11.9	0.0	28.4	80.9
20	2017/7/6 07:30	C_1	12.1	12.2	0.0	28.4	81.0

Source Code in R (2/6)

Create the dummy variable array_traffic 1

Create a new column, traffic_passing_by, in "data_array".

Assign all data cells to zero.

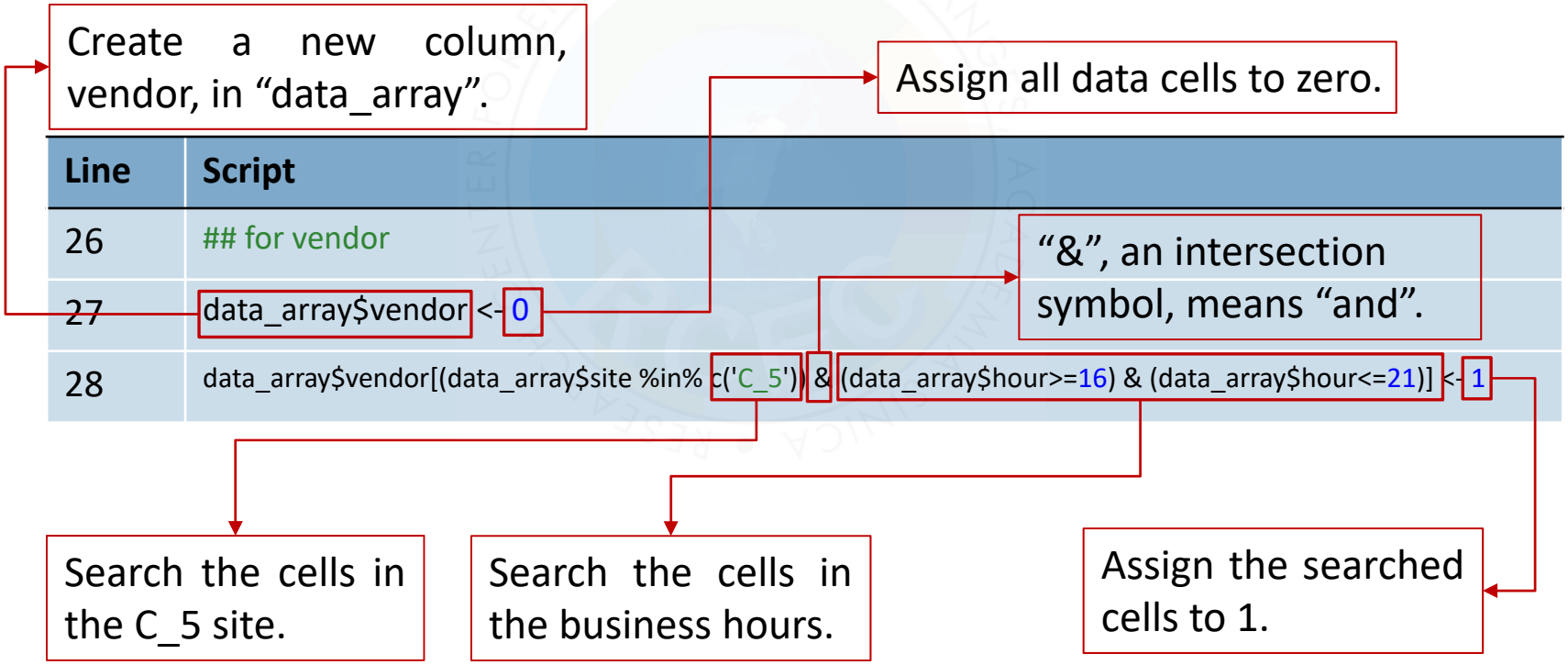
Line	Script
10	## for traffic type 1_traffic_passing_by
11	data_array$traffic_passing_by <- 0
12	data_array$traffic_passing_by[(data_array$site %in% c('C_1','C_2','C_3','C_4','C_5'))] <- 1

Assign the data cells of traffic_passing_by to 1 in those sites which the pollution source is nearby.

Please see page 4 for details.

Source Code in R (3/6)
Create the dummy variable array_vendor

- The business hours of this fried-chicken vendor was from 4-9 pm.



Line	Script
26	## for vendor
27	data_array$vendor <- 0
28	data_array$vendor[(data_array$site %in% c('C_5')) & (data_array$hour>=16) & (data_array$hour<=21)] <- 1

Source Code in R (4/6)
Multiple regression model

Assign a variable to store the result of the multiple regression analysis.

lm() is the function to establish the multiple regression model.

'formula=' is the parameter for the establishment of the regression model.

Line	Script
42	# build the multiple regression model
43	mlr<-lm(formula= site_pm2.5 ~ traffic_passing_by + traffic_stop_n_go + temple
44	+ market + gas_stat + vendor + school + season + high_level_pm2.5 + ws
45	+ temperature + rh, data=data_array)

Put y variable (site_pm2.5) of the regression model in the left side of '~' and x variables in the right side of '~'.

'data=' is used to assign the input data.

Input data include those data in page 6 from Line2 plus columns created in pages 8 & 9.

Source Code in R (5/6)
Multiple regression model_output

Use summary() to show the result of the regression analysis.

The result of the regression, “mlr”, is obtained in the previous page.

Use sink() to output the result in the ‘txt’ file.

Assign the file path of output data.

Line	Script
47	# show the result of the multiple regression model in the monitor
48	summary(mlr)
50	# save the result of the multiple regression model in the "txt" file, lines 51-53
51	sink("./output/mlr_result.txt")
52	summary(mlr)
53	sink()

The second sink() is to declare the end of the sink function.

The Result of the Regression Analysis

(example of "mlr_result.txt" mentioned in the previous page)

```
Residuals:
    Min      1Q  Median      3Q     Max
-39.172  -3.455  -0.862   2.476 198.137

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        26.237899   1.817143  14.439  < 2e-16 ***
traffic_passing_by  1.903888   0.223878   8.504  < 2e-16 ***
traffic_stop_n_go   3.572606   0.247489  14.435  < 2e-16 ***
temple              2.486907   0.202140  12.303  < 2e-16 ***
market              4.563443   0.226578  20.141  < 2e-16 ***
gas_stat            1.715894   0.222414   7.715 1.28e-14 ***
vendor              1.769434   0.328873   5.380 7.53e-08 ***
school             -0.957801   0.223850  -4.279 1.89e-05 ***
season             -0.611226   0.426443  -1.433    0.152
high_level_pm2.5    1.032735   0.007402 139.515  < 2e-16 ***
ws                 -1.025777   0.130087  -7.885 3.32e-15 ***
temperature        -0.602049   0.038890 -15.481  < 2e-16 ***
rh                 -0.142686   0.009456 -15.089  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.306 on 17155 degrees of freedom
  (552 observations deleted due to missingness)
Multiple R-squared:  0.7907,    Adjusted R-squared:  0.7906
F-statistic:  5402 on 12 and 17155 DF,  p-value: < 2.2e-16
```

3. These coefficients showed the contributions of different sources, e.g., the contribution of traffic type 1 is 1.9 μg/m^3 $PM_{2.5}$ increases in 5-min resolutions.

4. The assessments of source contributions were statistically significant (p-value < 0.05) except the coefficient of "season".

2. These variables explained 79.1% of the variability of the y variable.

1. The model of the multiple regression was statistically significant (p-value < 0.05).

Source Code in R (6/6)
Output the input data frame

Use write.csv() to output the input data in the ‘csv’ file.

Assign the data which you want to output. In our case, the input data of the the regression is “data_array”.

Line	Script
47	# save the data which is used in the multiple regression model in the "csv" file
48	write.csv(data_array, file="./output/data_community_workshop_output.csv",row.names = FALSE)

‘file=’ is used for the path and directory of the output file.

‘row.names = FALSE’ means not to output the row number in the file.

Thank you for your attention!