Forecasting and time variability analysis of Ozone concentrations using nitrogen oxide and meteorological variables as predictors

This manuscript ([permalink]) was automatically generated from [uiceds/cee-492-term-project-fall-2022-hydrograds@6042cc9] on October 30, 2022.

Authors

- **Jiewen Luo**
  · Noomi-Luo
  Department of CEE, University of Illinois at Urbana&Champaign

- **Rourou Ji**
  · JadeJi
  Department of CEE, University of Illinois at Urbana&Champaign

- **Bernardo Burbano**
  · BernieJBA
  Department of CEE, University of Illinois at Urbana&Champaign

CEE 492 Final Project Selection

1. Dataset description:

The dataset used in this project is a `CSV` file about the air quality in northern Taiwan collected in 2015 [https://www.kaggle.com/datasets/nelsonchu/air-quality-in-northern-taiwan], which include air quality data and meteorological monitoring data for research and analysis, originally from Environmental Protection Administration, Executive Yuan, R.O.C. (Taiwan). There are 25 observation stations in total. Columns in this CSV file are the following:

1. Time - The first column is the observation time of 2015

2. Station - The second column is the station name, there are 25 observation stations, those stations are showing at the table 1.

Table 1: A table contain all stations in Taiwan.

		station		
Banqiao	Cailiao	Datong	Dayuan	Guanyin
Guting	Keelung	Longtan	Pingzhen	Sanchong
Shilin	Songshan	Tamsui	Taoyuan	Tucheng
Wanhua	Wanli	Xindian	Xinzhuang	Xizhi
Yangming	Yonghe	Zhongli	Zhongshan	Linkou

3. Items - From the third column to the last one

4. item - `unit` - description

- SO_2 - `ppb` - Sulfur dioxide
- CO - `ppm` - Carbon monoxide
- O_3 - `ppb` - ozone
- PM_{10} - `μg/m^3` - Particulate matter
- $PM_{2.5}$ - `μg/m^3` - Particulate matter
- NO_x - `ppb` - Nitrogen oxides
- NO - `ppb` - Nitric oxide
- NO_2 - `ppb` - Nitrogen dioxide
- THC - `ppm` - Total Hydrocarbons
- NMHC - `ppm` - Non-Methane Hydrocarbon
- CH4 - `ppm` - Methane
- UVB - `UVI` - Ultraviolet index
- AMB_TEMP - `Celsius` - Ambient air temperature
- RAINFALL - `mm`
- RH - `%` - Relative humidity
- WIND_SPEED - `m/sec` - The average of the last ten minutes per hour
- WIND_DIREC - `degrees` - The average of the last ten minutes per hour
- WS_HR - `m/sec` - The average of an hour
- WD_HR - `degrees` - The average of an hour
- PH_RAIN - `PH` - Acid rain
- RAIN_COND - `μS/cm` - Conductivity of acid rain

Proposal:

The purpose of this project is to predict O_3 concentrations using measurements of concentration of other pollutants and available meteorological measurements. Ozone might be formed when heat and sunlight cause chemical reactions between oxides of nitrogen (NO_x) and Volatile Organic Compounds (VOC), which are also known as Hydrocarbons. Therefore it could be hypothesized that using measurements of NO_x as an independent variable a model could be developed to predict O_3 concentrations. Additionally, meteorological variables such as air temperature, relative humidity(RH) and ultraviolet index (UVB - UVI) could be included as independent variables to assess their influence on temporal variability of ozone. As an additional step wind-related variables such as mean wind velocity and direction will be included to study their effect on temporal variability of ozone.

After the air quality data has been processed the strongest O_3 predictors will be determined using PCA. PCA could be used to identify the main axes of variance within the dataset and explore underlying correlations that exist in a set of variables. Variables that are highly correlated cluster together. Using PCA 2D figures per each pair of variables are not needed, instead all the variables could be visualized simultaneously. Differences on PC1 are more important than differences on PC2. After plotting PCA plots, a heatmap could also be plotted to check the results. As additional criteria to identify the strongest predictors a LSTM network (long short-term memory network) can be used since the data used is time dependent. The network should contain several LSTM layers and fully-connected layers. The output should contain the pollution concentration and will point out the weights assigned to each correlated criterion, the values of such weights should also indicate what the strongest predictors are. Once the strongest predictors have been identified, genetic programming will be used to develop the models to predict O_3 concentrations.

Exploratory Data Analysis:

	count	mean	std	min	25%	50%	75%	max
AMB_TEMP	8682	1.18E-15	1.00	-2.47	-0.89	0.15	0.68	2.25
NMHC	8619	-4.74E-15	1.00	-1.32	-0.62	-0.30	0.29	16.26
NO	8462	3.20E-15	1.00	-0.75	-0.50	-0.32	0.05	23.55
NO2	8462	1.56E-15	1.00	-1.94	-0.78	-0.10	0.57	5.48
NOx	8462	1.62E-15	1.00	-1.54	-0.69	-0.20	0.42	14.67
O3	8685	-8.78E-16	1.00	-1.35	-0.80	-0.07	0.54	6.06
RH	8684	-1.56E-15	1.00	-4.85	-0.75	0.17	0.76	2.35
UVB	8684	3.53E-15	1.00	-0.56	-0.56	-0.56	0.12	4.81
WD_Hour	8680	2.64E-16	1.00	-1.58	-0.78	-0.59	1.01	2.32
WIND_DIREC_10min	8682	6.96E-16	1.00	-1.56	-0.78	-0.58	1.01	2.31
WIND_SPEED_10min	8682	-6.21E-15	1.00	-1.43	-0.72	-0.18	0.53	7.92
WS_HR	8680	-2.83E-15	1.00	-1.67	-0.80	-0.12	0.56	7.86

Heatmap Hour month O3 & Independent variable

	Dependent Variable	Independent variable: Nitrogen based				Independent variable: Metereological variables			
Station	**O3**	**NMHC**	**NO**	**NO2**	**NOx**	**Temperature**	**RH**	**UVB**	**W Direction 10 min**
Banqiao	0.991	0.984	0.966	0.966	0.966	0.991	0.991	0.991	0.991
Cailiao	0.978	0.000	0.958	0.958	0.958	0.977	0.978	0.000	0.977
Dayuan	0.988	0.000	0.975	0.975	0.975	0.980	0.955	0.000	0.852
Guanyin	0.980	0.000	0.969	0.969	0.969	0.973	0.980	0.000	0.977
Guting	0.988	0.966	0.983	0.983	0.983	0.985	0.983	0.000	0.985
Keelung	0.987	0.970	0.977	0.977	0.977	0.977	0.980	0.000	0.986
Linkou	0.973	0.000	0.966	0.966	0.966	0.973	0.973	0.000	0.972
Longtan	0.982	0.000	0.961	0.961	0.961	0.982	0.982	0.000	0.982
Pingzhen	0.988	0.000	0.968	0.968	0.968	0.988	0.988	0.000	0.983
Shilin	0.983	0.000	0.957	0.957	0.957	0.983	0.983	0.000	0.983
Songshan	0.985	0.967	0.976	0.976	0.976	0.985	0.985	0.000	0.984
Tamsui	0.961	0.000	0.956	0.956	0.956	0.000	0.000	0.961	0.000
Taoyuan	0.983	0.958	0.976	0.976	0.976	0.982	0.982	0.976	0.982
Tucheng	0.981	0.974	0.965	0.965	0.965	0.980	0.981	0.000	0.980
Wanhua	0.986	0.000	0.979	0.979	0.979	0.986	0.986	0.000	0.986
Wanli	0.984	0.000	0.976	0.976	0.976	0.983	0.976	0.000	0.976
Xindian	0.985	0.000	0.979	0.979	0.979	0.985	0.982	0.000	0.985
Xinzhuang	0.978	0.000	0.970	0.970	0.970	0.978	0.978	0.000	0.978
Xizhi	0.963	0.000	0.954	0.954	0.954	0.962	0.962	0.000	0.962
Yangming	0.979	0.000	0.964	0.964	0.964	0.979	0.979	0.000	0.000
Yonghe	0.987	0.972	0.978	0.978	0.978	0.984	0.987	0.000	0.987
Zhongli	0.988	0.973	0.967	0.967	0.967	0.986	0.986	0.000	0.986
Zhongshan	0.986	0.976	0.976	0.976	0.976	0.986	0.986	0.000	0.986

Predictive Modeling

The independent variables were segmented in pollutants and meteorological measurements. In order to visualize how the measurements change throughout the year the values were average per month. Then the resulting values were standardized using their mean. Once the values were standardized they were plotted against time.

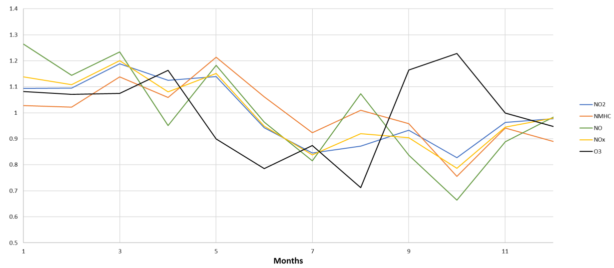

Figure 1: Standardized pollutants and ozone monthly concentration changes

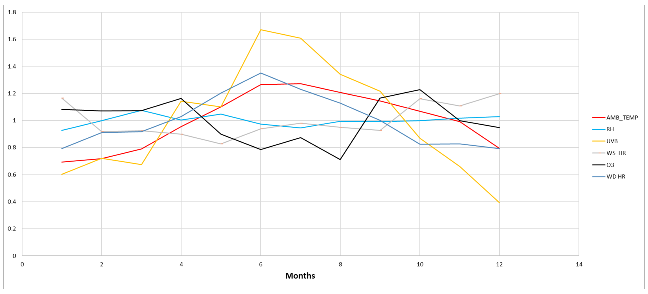

Figure 2: Standardized meteorological measurements and standardized ozone monthly concentration changes

As seen in the previous figure, O3 peaked in the months when concentration of the nitrogen based pollutants and non-methane hydrocarbons dropped. This is especially the case for NO concentrations (green line). This pattern of corresponding decreasing pollutant concentrations and increasing ozone could suggest that the pollutant concentrations are negatively correlated with ozone concentrations. This is also consistent with figure 1 (correlation plot)

In regards to the meteorological variables, UVB (ultraviolet index) and air temperature peak in the same months. Both temperature and UVB experience and increase in their values from the beginning of the year peaking in June. After June, both values experience a steady decrease. No discernable pattern can be observed in terms of the relation of the latter two variables and ozone concentrations.

Wind direction values are telling of changes in direction with respect to yearly average direction. The increase or decrease of the values shown in figure 3 correspond to a relative shift in direction of the wind compared to the yearly wind direction. These shifts in the direction of the wind can be used later on the forecasting of O3 concentration. Wind direction could help elucidate if O3 concentration from upwind neighboring locations could affect O3 values in the location of interest, Banquiao.

The exploratory data analysis suggests that in order to forecast ozone concentration the model inputs i.e. independent variables will have to be averaged over the month. Furthermore, such monthly mean values will have to be standardized using the mean annual corresponding values. Once the data has been standardized it will be used to train a model.

As a first iteration, a linear model with multiple independent variables will be optimized using available standardized measurements of ozone. A first model will be produced only using standardized pollutant values, and a second model will include as additional variables wind direction and upwind station standardized ozone concentrations from corresponding upwind stations. If the linear model mean square error, computed using predictions of ozone and observation, is below 0.5 a more involved model will be used. Two candidates for the second iteration of the predictive model will be considered. A fully connected neural network and a model produced with genetic programming packages in python for model discovery. In order to train the neural network pollutant measurements as well as from 9 out of the 12 months will be used as well as the corresponding mean wind direction values of Banquiao coupled with the ozone measurements from the 6 neighboring stations shown below. In order to validate the model, 12 the remaining data will be used.

Figure 3: Geographic location of 7 stations we focused on

Figure 3: Geographic location of 7 stations we focused on

This manuscript is a template (aka "rootstock") for Manubot, a tool for writing scholarly manuscripts. Use this template as a starting point for your manuscript.

The rest of this document is a full list of formatting elements/features supported by Manubot. Compare the input (`.md` files in the `/content` directory) to the output you see below.

Basic formatting

Bold text

Semi-bold text

Centered text

Right-aligned text

Italic text

Combined *italics and **bold***

~~Strikethrough~~

1. Ordered list item
2. Ordered list item
    a. Sub-item
    b. Sub-item
        i. Sub-sub-item
3. Ordered list item
    a. Sub-item

- List item
- List item
- List item

subscript: H_2O is a liquid

superscript: 2^{10} is 1024.

unicode superscripts^{0123456789}

unicode subscripts$_{0123456789}$

A long paragraph of text. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Putting each sentence on its own line has numerous benefits with regard to editing and version control.

Line break without starting a new paragraph by putting
two spaces at end of line.

Document organization

Document section headings:

Heading 1

Heading 2

Heading 3

Heading 4

Heading 5

Heading 6

A heading centered on its own printed page

Horizontal rule:

`Heading 1`'s are recommended to be reserved for the title of the manuscript.

`Heading 2`'s are recommended for broad sections such as *Abstract*, *Methods*, *Conclusion*, etc.

`Heading 3`'s and `Heading 4`'s are recommended for sub-sections.

Links

Bare URL link: https://manubot.org

Long link with lots of words and stuff and junk and bleep and blah and stuff and other stuff and more stuff yeah

Link with text

Link with hover text

Link by reference

Citations

Citation by DOI [1].

Citation by PubMed Central ID [2].

Citation by PubMed ID [3].

Citation by Wikidata ID [4].

Citation by ISBN [5].

Citation by URL [6].

Citation by alias [7].

Multiple citations can be put inside the same set of brackets [1,5,7]. Manubot plugins provide easier, more convenient visualization of and navigation between citations [2,3,7,8].

Citation tags (i.e. aliases) can be defined in their own paragraphs using Markdown's reference link syntax:

Referencing figures, tables, equations

Figure 4

Figure 5

Figure 6

Figure 7

Table 2

Equation 1

Equation 2

Quotes and code

> Quoted text

> Quoted block of text
>
> Two roads diverged in a wood, and I—
> I took the one less traveled by,
> And that has made all the difference.

Code `in the middle` of normal text, aka `inline code`.

Code block with Python syntax highlighting:

```python
from manubot.cite.doi import expand_short_doi

def test_expand_short_doi():
    doi = expand_short_doi("10/c3bp")
    # a string too long to fit within page:
    assert doi == "10.25313/2524-2695-2018-3-vliyanie-enhansera-copia-i-insulyatora-gypsy-na-sintez-ernk-
        modifikatsii-hromatina-i-svyazyvanie-insulyatornyh-belkov-vtransfetsirovannyh-geneticheskih-konstruktsiyah"
```

Code block with no syntax highlighting:

```
Exporting HTML manuscript
Exporting DOCX manuscript
Exporting PDF manuscript
```

Figures

Figure 4: A square image at actual size and with a bottom caption. Loaded from the latest version of image on GitHub.

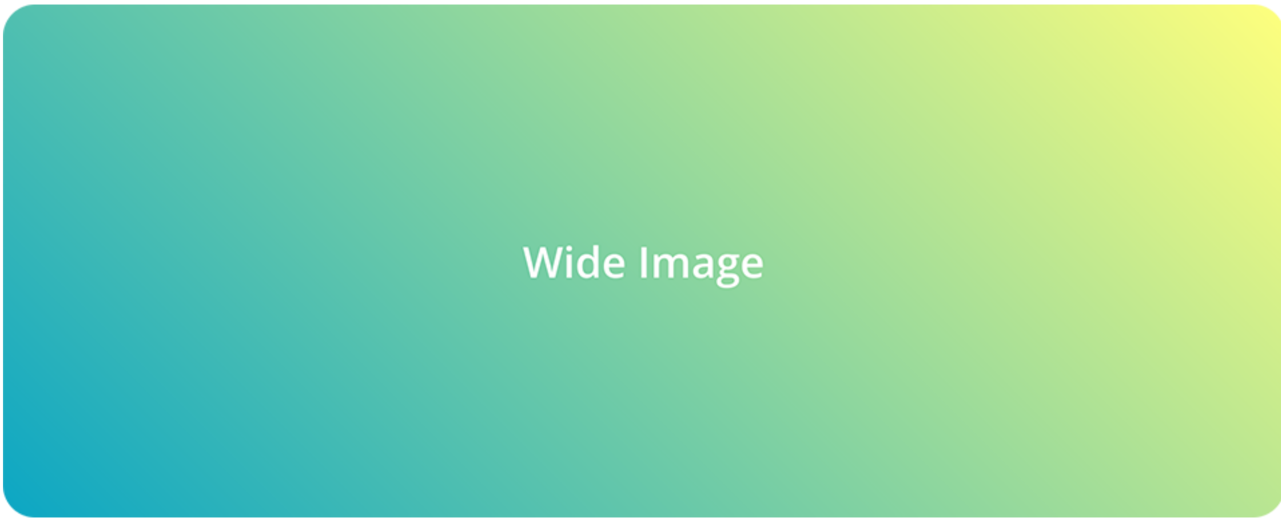

Figure 5: An image too wide to fit within page at full size. Loaded from a specific (hashed) version of the image on GitHub.

Figure 6: A tall image with a specified height. Loaded from a specific (hashed) version of the image on GitHub.

Figure 7: A vector `.svg` image loaded from GitHub. The parameter `sanitize=true` is necessary to properly load SVGs hosted via GitHub URLs. White background specified to serve as a backdrop for transparent sections of the image.

Tables

Table 2: A table with a top caption and specified relative column widths.

Bowling Scores	**Jane**	**John**	**Alice**	**Bob**
Game 1	150	187	210	105
Game 2	98	202	197	102
Game 3	123	180	238	134

Table 3: A table too wide to fit within page.

	Digits 1-33	**Digits 34-66**	**Digits 67-99**	**Ref.**
pi	3.14159265358979323846264338327950	288419716939937510582097494459230	781640628620899862803482534211706	`piday.org`
e	2.71828182845904523536028747135266	249775724709369995957496696762772	407663035354759457138217852516642	`nasa.gov`

Table 4: A table with merged cells using the `attributes` plugin.

	Colors	
Size	**Text Color**	**Background Color**
big	blue	orange
small	black	white

Equations

A LaTeX equation:

$$\int_0^\infty e^{-x^2} dx = \frac{\sqrt{\pi}}{2} \tag{1}$$

An equation too long to fit within page:

$$\begin{split} x = a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t \\ + u + v + w + x + y + z + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 \end{split} \tag{2}$$

Special

⚠ WARNING *The following features are only supported and intended for `.html` and `.pdf` exports. Journals are not likely to support them, and they may not display correctly when converted to other formats such as `.docx`.*

LINK STYLED AS A BUTTON

Adding arbitrary HTML attributes to an element using Pandoc's attribute syntax:

> Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Adding arbitrary HTML attributes to an element with the Manubot `attributes` plugin (more flexible than Pandoc's method in terms of which elements you can add attributes to):

> Manubot Manubot Manubot Manubot Manubot. Manubot Manubot Manubot Manubot. Manubot Manubot Manubot. Manubot Manubot. Manubot.

Available background colors for text, images, code, banners, etc:

`white` `lightgrey` `grey` `darkgrey` `black` `lightred` `lightyellow` `lightgreen` `lightblue` `lightpurple` `red` `orange` `yellow` `green` `blue` `purple`

Using the [Font Awesome](https://fontawesome.com/) icon set:

✔ ? ★ 🔔 ⊗ ⋯

Light Grey Banner
useful for *general information* - manubot.org

Blue Banner
useful for *important information* - manubot.org

Light Red Banner
useful for *warnings* - manubot.org

References

1. **Sci-Hub provides access to nearly all scholarly literature**
   Daniel S Himmelstein, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, Casey S Greene
   eLife (2018-03-01) https://doi.org/ckcj
   DOI: 10.7554/elife.32822 · PMID: 29424689 · PMCID: PMC5832410

2. **Reproducibility of computational workflows is automated using continuous analysis**
   Brett K Beaulieu-Jones, Casey S Greene
   Nature biotechnology (2017-04) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6103790/
   DOI: 10.1038/nbt.3780 · PMID: 28288103 · PMCID: PMC6103790

3. **Bitcoin for the biological literature.**
   Douglas Heaven
   Nature (2019-02) https://www.ncbi.nlm.nih.gov/pubmed/30718888
   DOI: 10.1038/d41586-019-00447-9 · PMID: 30718888

4. **Plan S: Accelerating the transition to full and immediate Open Access to scientific publications**
   cOAlition S
   (2018-09-04) https://www.wikidata.org/wiki/Q56458321

5. **Open access**
   Peter Suber
   MIT Press (2012)
   ISBN: 9780262517638

6. **Open collaborative writing with Manubot**
   Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter
   Manubot (2020-05-25) https://greenelab.github.io/meta-review/

7. **Opportunities and obstacles for deep learning in biology and medicine**
   Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, … Casey S Greene
   Journal of The Royal Society Interface (2018-04) https://doi.org/gddkhn
   DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

8. **Open collaborative writing with Manubot**
   Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter
   PLOS Computational Biology (2019-06-24) https://doi.org/c7np
   DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653