An Application of Statistical Learning on Understanding the Effect of Meteorology and Air Pollution to Visibility Reduction in Hong Kong International Airport

Kaipeng Zhong

Department of Geography and Environmental Engineering

Johns Hopkins University

Term Project for

EN.570.608.01.FA14

Data Analytics for Engineering, Policy Analysis and Management

Abstract

The reduced visibility in major metropolitans of China displays an increasing trend over the past 30 years. The visibility problems have raised public concern on its indication to deteriorated air quality. Some key pollutants such as fine particulates, ozone and nitrogen oxides and meteorological conditions such as mean temperature, relative humidity and wind were selected as covariates in the models. Past research to Hong Kong studied the contribution of air pollutants to reduced visibility using classical regression approaches. The pollution pattern of Southern China might have changed in the past ten years as polluted industries were moved north and the number of vehicle grows dramatically. However, there is few paper published with the latest data in Southern China. In this paper, eight statistical learning techniques were adopted to quantify the relative importance of pollutant concentration and meteorological conditions regarding to visibility reduction in recent years using data obtained from Hong Kong Environmental Protection Department. We concluded from the results that fine particulate, relative humidity and mean temperature are determinant factors.

Keywords: *visibility, classification, gam, bayesian, ann*

An Application of Statistical Learning on Understanding the Effect of Meteorology and Air Pollution to Visibility Reduction in Hong Kong International Airport

The reduction of visibility in metropolitans of China has become more severe in recent years which raises public concern on its relation to deteriorated air quality. It has been confirmed by study that the air pollution episodes were related to increased energy consumption (Chan & Yao, 2008). Historical observation from Hong Kong Observatory has shown an increased trend of days with reduced visibility since 1968. The study investigated the effect of seasonality and climate on visibility with 15-year data (Chang & Koo, 1986). The reduction of visibility not only affect the aesthetic view of a city but also causing danger to air and ground traffic. Apart from meteorological phenomena such as mist, fog and precipitation, the reduction of visibility also associated with suspended particulates and light-absorbing gaseous particles. The presence of air pollutants only explained part of the story. The reduced visibility in urban Hong Kong is far more complicated, included the combination of many variables acting against each other (Chang & Koo, 1986). Meteorological conditions such as relative humidity and mixing height were reported to have great influence on visibility (Qu, Wang, Zhang, Wang, & Sheng, 2015).

In real world, relationships between variables are often nonlinear and thus prediction from a simple linear model is not ideal, especially when studying a complex physical and chemical system such as the atmosphere. The relationship between visibility and pollutant concentration has suggested to be nonlinear and more specifically multiplicative (Butcher & Charlson, 1972). By using the Generalized Additive Model (GAM) or Multi, the problem of nonlinear relationship can be solved as each term in a GAM is a nonlinear function with certain degrees of freedom. The nonlinear function is estimated by scatterplot smoother (James, Witten, Hastie, & Tibshirani, 2013).

Two types of classification model were used in this study. The first type is likelihood-based models such as linear or nonlinear regression models. To make them applicable to binary response variables, the inverse of logistic function can convert them into continuous log odds so that a linear or nonlinear function can fit. The second type is tree-based model developed by Breiman. While logistic regression is technically a linear classification model based on maximum likelihood estimation, Classification and Regression Tree (Breiman L. , 1984), Random Forest (Breiman L. , 2001) and Bayesian Classification and Regression Tree (Chipman, George, & McCulloch, 1998) are based on recursive partitioning of decision tree. Artificial Neural Networks (ANN) is an adaptive learning model consists of an input layer, one or more hidden layers and an output layer. Each node in hidden layers represent a simple function, normally a logistic function.

By building the models, we hope to be able to explain research questions such as 1) Does the reduction in visibility more related to air pollution, climate change or both? 2) What conditions favor the reduction of visibility in the recent years? Are they similar to past research? To address these questions, we analyzed recent 2-year ground observation data for weather and air quality from Hong Kong International Airport. The result from this study may not be fully relied on to unveil underlying factors responsible for physical phenomena due to limited spatial and temporal range of study. However, these findings can be utilized to provide preliminary guidance for further field experiments and numerical simulations.

Background

Geography, Climate and Air Quality of Hong Kong

Hong Kong is located at the south-eastern tip of China and it consists of Hong Kong Island, Lantau Island, the Kowloon Peninsula, the New Territories and 262 outlying islands. It is the most populated city in Southern China with over 72 million of people on 1,104 square kilometers of land. (GovHK, 2014) Up to September of 2014, the total registered vehicles in Hong Kong is about 763,000. (Census and Statistics Department, 2014)

The type of climate in Hong Kong is Cwa according to the Köppen-Geiger Climate Classification, which indicates warm temperate climate with dry winter and hot summer. This is in accordance to the fact published by Hong Kong Observatory (HKO). The number of instances of reduced visibility reached a record high in 2004 and became very noticeable to Hong Kong citizens who perceived the phenomenon as related to increased air pollution.

According to HKO, the reduced visibility typically occurred in winter and spring was association with wind. During this time, northerly surges of the monsoon was weak and the effect of northeast monsoon Hong Kong was also relatively weak. In summer and early autumn, episodes of reduced visibility often occur associated with northerly or northwesterly winds with low boundary layer height due to tropical cyclones that were a few hundred kilometers to the east of Hong Kong. In short, it is inferred that the deteriorating visibility is associated with an increase in the strength of sources of suspended particulates inland to the north of Hong Kong. (Lam & Lau, 2005)

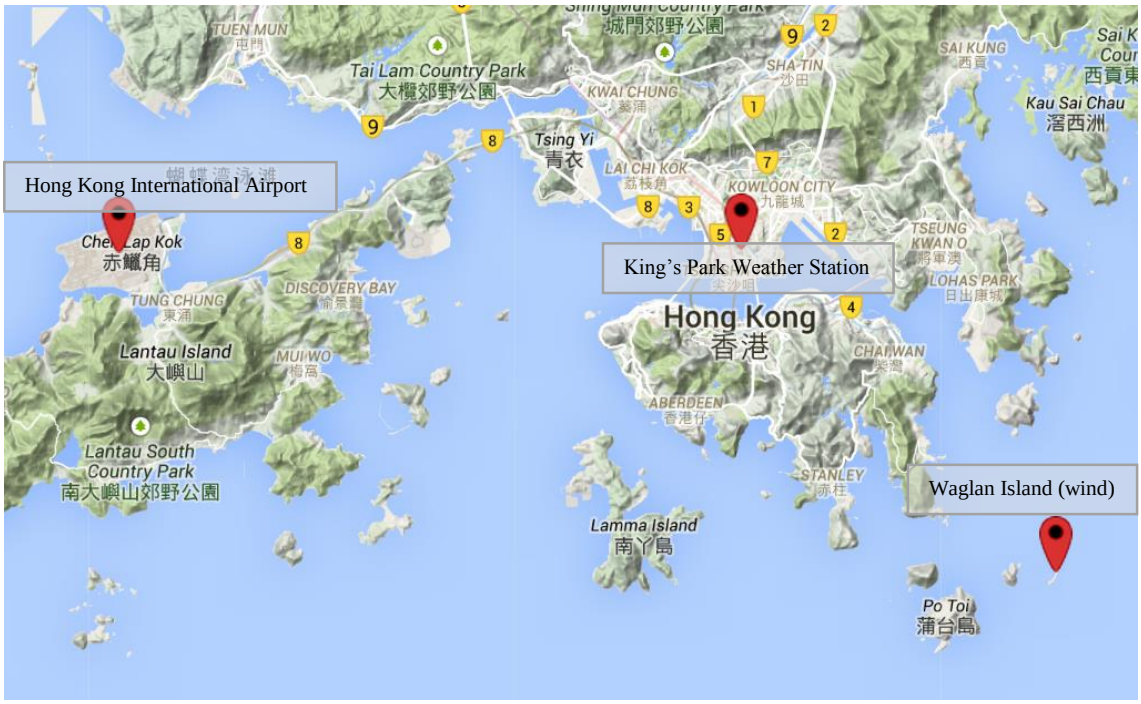

Figure 1. Overview of air monitoring and weather stations in this study.

Figure 1 shows the aerial view of Hong Kong and the locations of three study sites. The visibility data was obtained at Hong Kong International Airport, meteorological data was obtained at Hong Kong Observatory at King's Park and background wind data was obtained at Waglan Island at the southeastern tip of Hong Kong. The pollution data was a spatial average of all the ground monitoring station in Hong Kong.

Previous Research

The past studies from Hong Kong Observatory addressed the effect of atmospheric suspended particulate on reduced visibility (visibility below 8 km) by excluding cases associated with mist, fog, rain or relative humidity ≥ 95%. (Leung, Wu, & Yeung, 2008) There was also studies conducting wind attribution analysis on visibility using linear regression with 12 wind sector centered at 0° 30° 60° … 300°, 330° and trajectory analysis on suspended particulates using the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model from

NOAA. (Leung & Lam, 2008). A study from the Air Quality Group of UC Davis in 1978 addressed the effect of particulate aerosols and meteorological conditions simultaneously using a multiplicative model rather than a linear model. The first two studies from Hong Kong focused on only single variable, which may not accurately reflect the relative importance of air pollution and meteorological conditions. The study from UC Davis has successfully combined these two but the result might not applicable to Southern China. In fact, wind direction is also an important consideration as regional transportation of atmospheric suspended particulates exists.

Data Description

Air quality and meteorology data used in this study was collected from Hong Kong Environmental Protection Department and Hong Kong Observatory. The air quality data is hourly and converted to daily by time-averaging to merge with the meteorology data by date. The dataset covers the period from March 8, 2012 to December 29, 2013 and all the missing data were removed. The processed dataset contains 660 observations, 17 predictive variables and 1 response variable – Reduced Visibility. This variable is binary and has two levels – "yes" and "no". The response variable contains 436 days with no reduction and 224 days with reduction in visibility. The visibility readings were measured in an average of 10-minute period before the clock hour of the visibility meter near the middle of the south runway at the Hong Kong International Airport. Reduced Visibility in this data set is represented as number of hours that reduced visibility occurred. According to the definition from Hong Kong Observatory, Reduced Visibility refers to visibility below 8 kilometer in the absence of fog, mist or precipitation. Predictors from air quality data set have units of microgram per cubic meter ($\mu g / m^3$) for

particulate matter and parts per million (ppm) for gaseous species such as ozone, nitrogen oxides, sulfate dioxide and carbon monoxide.

Method

Two types of model, regression-based model and classification-based model were applied in this study. Regression-based models include GLM, GAM, MARS and classification-based models include SVM, CART, Random Forest, BART and Artificial Neural Networks (ANN). Logistic regression was used for the purpose of classification. The goodness-of-fit and classification accuracy were compared. The goodness-of-fit was evaluated using the original set of data that train the model. The classification accuracy was calculated by Leave-One-Out-Cross-Validation (LOOCV) that each time one observation is predicted using the rest of the data and the classification accuracy is calculated by the number of correct classification over the total number of observation. The LOOCV was also applied to parameter selection for choosing a model with the best performance.

Data Transformation

The original data set contains variables that are better to convert to categorical form because of incompleteness and incompatibility. The variable *rainfall* in the original data set was in mixed format containing numeric values and characters. No rain is denoted by hyphen "-", total daily rainfall less than 0.05mm is denoted by word "trace" and actual values for the rest. To standardize the *Rainfall* variable, it was converted to categorical in accordance to the definition of rain by China Meteorological Administration (CMA).

Table 1. Classification of rain level modified from China Meteorological Administration (CMC) with new class – Trace.

Level of Rainfall	24-h Average (mm)	12-h Average (mm)
0 – No	0	0
1 – Trace	< 0.05	–
2 – Light	< 10	< 5
3 – Moderate	10-24.9	5-14.9
4 – High	25-49.9	15-29.9
5 – Heavy	50.99.9	30-69.9
6 – Severe	100-249.9	70-139.9
7 – Extreme	> 250	> 140

The variable *winddir* in the original data set was recorded in degree. The polar coordinate system used in representing direction is not compatible with the Cartesian coordinate system used in linear or nonlinear function in statistical models. For example, 359° and 1° are both classified as northerly wind and there are only 2° difference in between, but they are considered dramatically different in Cartesian coordinate system. To overcome the difference in measurement system, the *winddir* variable was converted to categorical variable with 12 wind sector centered at 0° 30° 60° … 300°, 330° as shown in *Figure 2*. Finally, the response variable *reducedvis* was converted to binary representing whether reduced visibility occurred in anytime of a day.

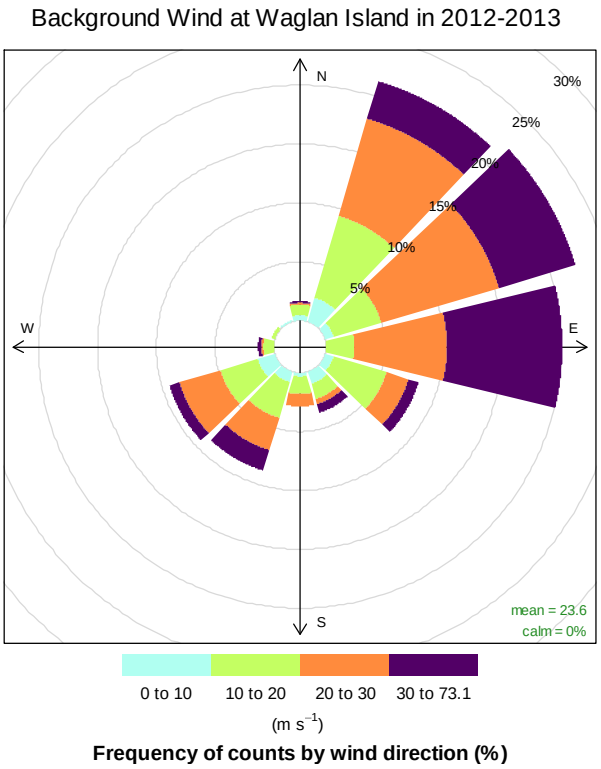

Figure 2. Hong Kong background wind frequency map in 2012 and 2013 with 12 wind direction sectors. Frequency count is based on percentage. Wind speed is shown in color bins with the unit of m/s.

Model Comparison and Selection

Within the same model, Leave-One-Out-Cross-Validation (LOOCV) was used to calculate out-of-sample classification accuracy and models with the highest accuracy was selected to compare with other best-selected models. As the name suggests, Leave-One-Out Cross-Validation (LOOCV) use one observation for validation and the remaining observations for training the model. This is repeated until every observation has a predicted value. (James, Witten, Hastie, & Tibshirani, 2013)

In binary classification, the performance of a model classifier is usually measured by a 2x2 confusion matrix. The term “sensitivity” calculates the true negative rate. In this study, it calculates the rate that the model correctly classified a day with no reduction in visibility. The term “specificity” calculates the true positive rate, which is the rate that the model correctly classified a day with reduction in visibility. Models were not only compared by classification accuracy, but also their performance in accurately identified reduced visibility

Regression Models. Variable selection for GLM was accomplished by stepwise and exhaustive screening algorithm based on Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC). For likelihood-based model such as GLM and GAM, 5 models with the lowest AIC and 5 model with the lowest BIC chosen by automated variable selection were used as model candidates to LOOCV. Since Generalized Additive Model (GAM) is an extension to GLM that allows some degrees of freedom for the predictive variables to be nonlinear, variables used to construct the GAM were inherited from exhaustive screening results from GLM and stepwise AIC algorithm for GAM. Multivariate Adaptive Regression Splines (MARS) is an automatic process that does variable selection itself. The controlling parameter is a penalty term that determine how close a spline can follow the data points.

Classification Models. The controlling parameter for a single tree model is the tree size, i.e. the number of split. Cross validation was performed to select size that yields the lowest deviance. For models with ensemble-of-trees such as RF and BART, the controlling parameters are the penalty terms *k*. The out-of-sample classification accuracy was used as the criterion for model selection. This was accomplished by a FOR Loop in R script to try different values of *k* until the maximum accuracy was found. A single hidden layer ANN was applied in this study.

Similar to penalty factor *k*, the number of nodes in hidden layer and decay factor were tuned to get the optimum classification accuracy.

Result and Discussion

In this section, the classification accuracy from LOOCV and fitted value of the best-performing models was calculated by confusion matrix. Variable importance and influence will be discussed at the end of this section with separation of variable-selection-based models and variable-importance-based models.

Table 2. Classification statistics for eight best-performing models. Overall accuracy was used as the selection criterion.

Models	LOOCV			Fitted Value
	Sensitivity	Specificity	Overall Acc.	Accuracy
GAM	0.908	0.777	0.864	0.885
MARS	0.890	0.760	0.848	0.862
ANN	0.862	0.759	0.827	0.873
CART	0.853	0.746	0.817	0.886
GLM	0.906	0.732	0.847	0.850
SVM	0.899	0.719	0.838	0.858
RF	0.908	0.705	0.840	0.827
BART	0.894	0.696	0.830	0.868
Mean	0.890	0.737	0.839	0.864
SD	0.020	0.027	0.014	0.018

Classification accuracy

Table 2 summarizes the classification accuracy obtained from LOOCV and fitted values of the best-performing models. Sensitivity and specificity were calculated from LOOCV and classification accuracy was calculated for both LOOCV and model fitted values. All models seem to have a biased classification that sensitivity is much higher than specificity. The mean classification accuracy for eight best-performing models is 0.839. In correctly identifying days

with reduced visibillity, the mean accuracy (specificity) is 0.737. Since the goal of our study is to understand potential factors that result in reduced visibility, we have more concern on model's ability to classify days with reduced visibility. Therefore, *Table 2* was sorted in descending order of specificity. Among all models, GAM has the highest accuracy and specificity. The second and the third highest specificity are MARS and ANN. Models like GLM, SVM and RF have high sensitivity, but their specificity are relatively low. Though the goodness-of-fit for RF is the lowest among other models, its overall classification accuracy is still above the mean.

Table 3. Normalized variable importance for eight best-performing models by *caret*

Variables	GLM	GAM	MARS	SVM	CART	BART	RF	ANN	TOTAL
pm25	1.000	1.000	1.000	1.000	1.000	0.819	1.000	1.000	7.819
pm10	-	-	0.127	0.965	0.753	1.000	0.736	0.579	4.160
co	-	0.357	-	0.955	0.827	0.468	0.624	0.619	3.849
relhum	0.842	-	0.517	0.627	0.301	0.436	0.400	0.513	3.636
no2	0.222	-	-	0.943	0.695	0.496	0.506	0.553	3.415
meantemp	0.321	0.567	0.217	0.836	0.046	0.379	0.296	0.505	3.166
so2	-	0.710	0.077	0.841	0.157	0.554	0.297	0.480	3.116
dewpointtemp	-	0.594	-	0.836	0.382	0.362	0.355	0.301	2.831
windspeed	0.140	0.710	-	0.636	0.095	0.371	0.257	0.283	2.494
pressure	0.189	-	-	0.796	0.068	0.199	0.267	0.882	2.400
cloudpct	-	-	0.171	0.572	0.095	0.724	0.200	0.570	2.331
radiation	-	-	-	0.721	0.320	0.470	0.295	0.460	2.265
rainfall	0.222	0.225	-	0.701	0.000	0.320	0.073	0.429	1.970
o3	-	0.044	0.113	0.601	0.216	0.195	0.263	0.496	1.928
evaporation	-	-	-	0.715	0.000	0.434	0.191	0.572	1.912
sunshine	-	-	-	0.642	0.155	0.111	0.184	0.368	1.462
winddir	-	-	-	0.572	0.016	0.068	0.171	0.195	1.022

Variable importance and influence

The result of variable selection and normalized variable importance was summarized in *Table 3*. The calculation of variable importance was facilitated by *caret* package of R (Kuhn, 2008). Almost all models agreed PM2.5 (*pm25*) the most important factor in reduced visibility and it is almost twice higher than the second highest factor PM10. They can be grouped under

one category named particulate. The most important meteorological factor is relative humidity and the second most important meteorological factor is mean temperature. Our finding for the positive effect of particulate and relative humidity were in accordance to past research (Lee & Sequeira, 2001) (Chan & Yao, 2008). Another noticeable predictors are carbon monoxide (*CO*) and nitrogen dioxide (*NO2*). Together with particulate, these three air pollutants are highly related to vehicle emission, which is now a common problem in China metropolitans.

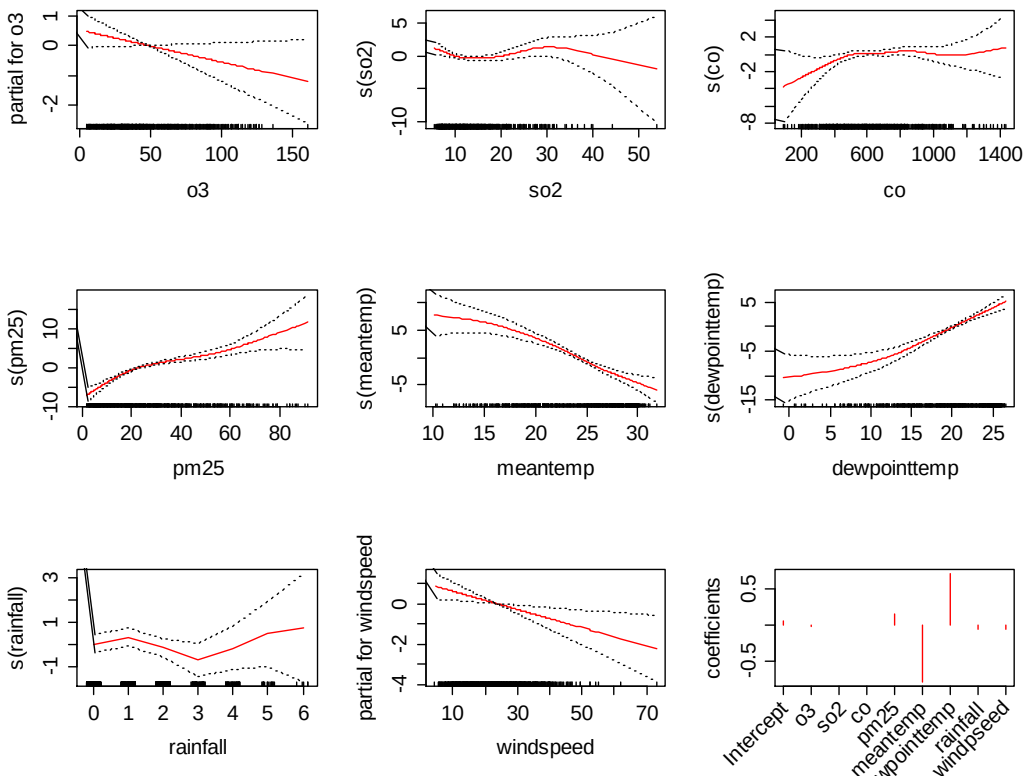

Figure 3. Plot on main-effect functions in GAM and partial dependence on linear functions.

To obtain a more sensible result for the effect of each variable, the model with the highest classification accuracy was discussed. The stepwise algorithm for GAM with spline smoother

selected 8 out of 17 predictors. Ozone and wind speed were fitted with linear function and the other variables were fitted with smoothing splines in 3 degrees of freedom. As shown in lower right of Figure [3], visibility is significantly influenced by the level of PM2.5 and temperature. The positive coefficients for PM2.5 and dew point temperature indicate reduced visibility is positive correlated to them. The effect of PM2.5 on visibility is intuitive as it scatters transmitted light by a process called Mie Scattering. This process is independent of wavelength and thus resulted in grey sky (Mishchenko, Travis, & Lacis, 2002). The dew point temperature is closely related to humidity of the air. High dew point temperature indicates high humidity and *vice versa*. The high level of water vapor favors the formation of fog, mist and haze that reduce the visibility. The mean temperature can be viewed as a surrogate for time and seasons. Past study pointed out that reduced visibility was more common in spring than summer when the lower atmosphere is stable (Chang & Koo, 1986).

Conclusion

This study investigates the effect of common air pollutants and meteorological conditions to visibility using eight models with different classification algorithm. Eight models were developed using ground observation data with satisfied classification accuracy of 0.84 in average. The model comparison result shows that Generalized Additive Model (GAM) has the highest classification accuracy for this data set. It indicates that the effect air pollutants and meteorological condition to visibility was additive. Through studying the variable importance for each model, we are able to answer some of the research questions raised up in the beginning of this report. We found that particulate (PM10, PM2.5) and source emission (CO, NO2, SO2) are important to the reduction in visibility in recent two years. However, the result from these

models also points out the importance of humidity (relative humidity, dew point temperature) and daily mean temperature. The partial dependence plot for GAM suggest that the contribution from these two meteorological factors are in fact greater than particulate and other gaseous compounds despite they have lower importance. A possible explanation to this paradox is their good representation of time and seasonality as Hong Kong has distinctive dry winter and humid summer. Also, all variables in building the models are time-dependent that may be self-correlated. If true, the underlying assumption of regression modeling that each observation is independent would be violated. In that case, detrended data or models addressed correlated data such as Generalized Estimated Equation (GEE) can be used.

References

Breiman, L. (1984). *Classification and Regression Trees.* Chapman&Hall/CRC.

Breiman, L. (2001). Random Forests. *Machine Learning*(45), 5-32. doi:10.1023/A:1010933404324

Butcher, S., & Charlson, R. (1972). *An Introduction to Air Chemistry.* New York: Academic Press.

Census and Statistics Department. (2014, October 1). Hong Kong Monthly Digest of Statistics. Hong Kong Island, Hong Kong SAR. Retrieved from http://www.statistics.gov.hk/pub/B10100022014MM10B0100.pdf

Chan, C., & Yao, X. (2008). Air pollution in mega cities in China. *Atmos. Environ.*, 1-42.

Chang, W., & Koo, E. (1986). A study of visibility trends in Hong Kong. *Atmospheric Environment*, 1847-1858.

Chipman, H., George, E., & McCulloch, R. (1998). Bayesian CART model search. *J Am Stat Assoc*(93), 935–948.

GovHK. (2014, November 14). *Hong Kong - the Facts.* Retrieved from GovHK: http://www.gov.hk/en/about/abouthk/facts.htm

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. Software*(28), 1–26.

Lam, C., & Lau, K. (2005). *Scientific Background of Haze and Air Pollution in Hong Kong.* Hong Kong: The 13th Annual Conference of Hong Kong Institution of Science.

Lee, Y. L., & Sequeira, R. (2001). Visibility degradation across Hong Kong: its components and their relative contributions. *Atmospheric Environment*, 5861–5872.

Leung, Y., & Lam, C. (2008). Visibility Impairment in Hong Kong – A Wind Attribution Analysis. *Bulletin of Hong Kong Meteorological Society*, 33-48.

Leung, Y., Wu, M., & Yeung, K. (2008). A Study on the Relationship among Visibility, Atmospheric Suspended Particulate Concentration and Meteorological Conditions in Hong Kong. *Acta Meteorologica Sinica*, 461-469.

Mishchenko, M., Travis, L., & Lacis, A. (2002). *Scattering, Absorption, and Emission of Light by Small Particles.* . New York: Cambridge University Press.

Qu, W., Wang, J., Zhang, X., Wang, D., & Sheng, L. (2015). Influence of relative humidity on aerosol composition: Impacts on light extinction and visibility impairment at two sites in coastal area of China. *Atmospheric Research*, 500-511.

Footnotes

[1][Add footnotes, if any, on their own page following references. For APA formatting requirements, it's easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. *(Note: If you delete this sample footnote, don't forget to delete its in-text reference as well. That's at the end of the sample Heading 2 paragraph on the first page of body content in this template.)*]

Tables

Table 1

[Table Title]

Column Head	Column Head	Column Head	Column Head	Column Head
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

Figures

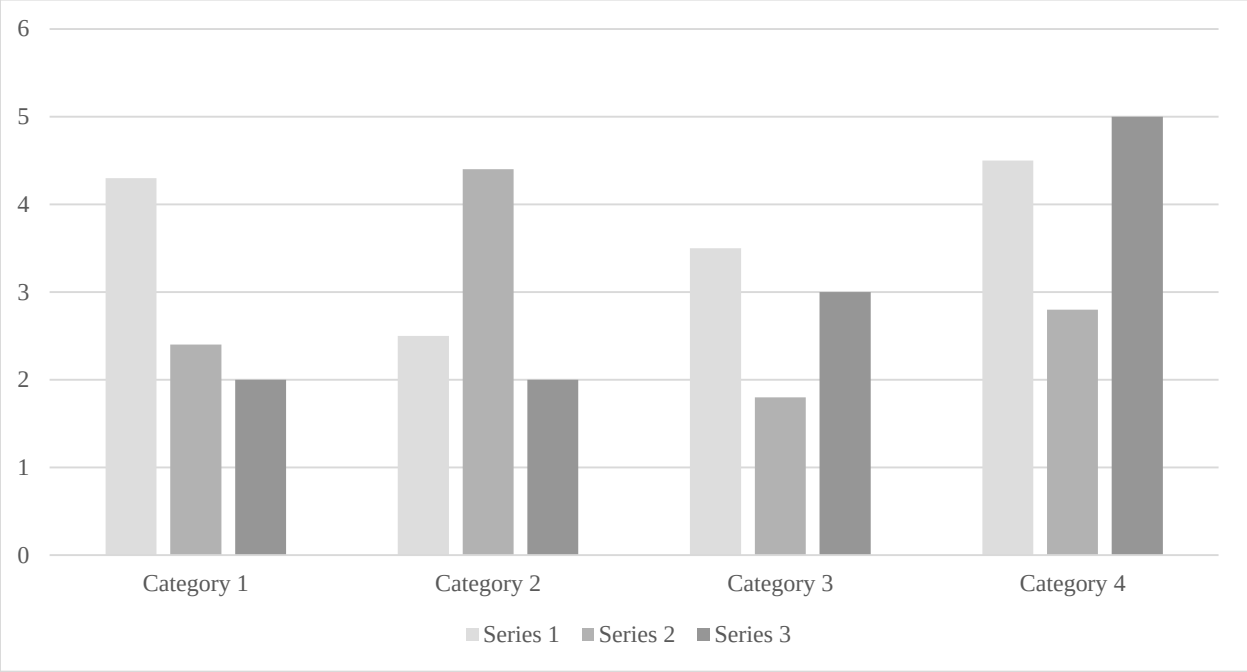

Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the *APA Style Manual, 6th Edition*.