IDENTIFYING AND QUANTIFYING THE IMPACT OF AIR POLLUTION SOURCE AREAS BY NONPARAMETRIC TRAJECTORY ANALYSIS

by

Chien-Cheng Pan

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ENVIRONMENTAL ENGINEERING)

December 2008

Dedication

To my parents and my wife, for their love, support, guidance, and encouragement: my parents whose greatest wishes have been their children's highest education, and the love of my wife.

Acknowledgements

Getting my PhD was a long term waiting and expecting. There were several times I felt like it is an impossible mission. Fortunately, family' and friends' supports made me focus on my target and path. I would like to express my deepest appreciation to my advisor, Dr. Ronald Henry. Not only was my PhD completion impossible without his academic and financial support, but also his great sharing with other aspects of life.

I would also like to thank my dissertation committee Dr. George Chilingarian and Dr. Gareth James for their guidance and helpful suggestions.

I am forever indebted to my lovely wife, Hsiao-Fong, for her unconditional love and support, her daily encouragements, and her patiently sharing the hard time with me. I am also thankful to my friends in church for their supports and prayers for my family.

Above all, I thank and give glory to almighty God.

Table of Contents

List of Tables

List of Figures

Abstract

In order to improve air quality, it is necessary to identify and quantify the sources of airborne pollution. Local emissions are more easily to control compared to regional emissions since multiple agencies and states are not involved in the regulatory process.

Generally two types of air quality models, source – oriented models and receptor – oriented models, are used to evaluate the impact of emission on air quality on a local, regional, and global scale. Source – oriented models require detailed information on emission composition, rates and local meteorological data. Therefore, they are not suitable for sources of fugitive emissions and intermittent or temporary emissions, which cannot easily be quantified. On the other hand, receptor models need chemical composition data to identify and quantify sources affecting the monitoring sites. However, pollutants without distinguishable "fingerprints", such as SO_2, O_3 cannot be apportioned by this method.

A new hybrid source – receptor model was previously develop and is called Nonparametric Trajectory Analysis (NTA). It is based on nonparametric kernel smoothing and backtrajectory analysis. NTA was developed to identify and quantify local sources of species measured on a very short time scale, i.e, minute, and it has gotten some encouraging results. However, NTA sometimes produces artifacts areas that appear to be sources but not, this is especially true for sources very close to the receptor. A major objective of this study is to address this difficulty.

The NTA gives a map of the average concentration at the receptor when the air passes over each point on the map. This NTA map is obviously related to the local sources affecting the receptor, but it is not a map of the sources. One way to extend the NTA method is the Point Source Response (PSR) method. The NTA map can be considered a linear combination of responses to number of point sources. The NTA map for a point source at each point on a grid is calculated. Next, the weighted sum of the PRS maps that best fits the NTA map for the real data is estimated by principal components

regression. In this way, the size and location of source affecting the receptor are estimated.

This method is illustrated by application to 1- minute SO_2 data from Long Beach, and 1-minute PM_{10} data for Rubidoux along with meteorological data from nearby monitoring stations in South Coast Air Basin of Southern California. The result identified the Long Beach harbor and transportation hubs close to the intersection of freeway 710 and freeway 405 and Long Beach Airport as major SO_2 sources. For the Rubidoux area, aggregate, and asphalt factories, and construction sites are identified as source of PM_{10}.

Chapter 1

Introduction

A. General Introduction

In order to know whether National Ambient Air Quality Standards can be met or not, the government of the United States built up thousands of air quality monitors all over its territories. These monitors usually report 1-minute average data of various species of air pollutants, and collect meteorological data. These air pollutions based on their emission sources can be classified as global, regional or local scale sources. For environmental concerns, the local sources are much more important than the other sources. Not only because of local sources are close to our living surroundings, but also control of local sources is often easier and more cost effective than control of

regional sources since multiple agencies and states are not involved in the regulatory process.

To understand the impact of emission sources on the air quality, air quality models are required since this is the most effective way to identify and quantify the sources of pollution. Generally, they can be classified in two categories: source-oriented air quality models, receptor-oriented models. These models will go into details in the next chapter.

B. Statement of Problem

This study tends to follow Dr. Pazokifard's research (Pazokifard, 2007) to take advantage of Nonparametric Trajectory Analysis (NTA). NTA is a new hybrid source-receptor model, which seeks to takes advantages of 1- minute observed air pollution concentrations and meteorological data to identify and quantify the local sources of pollution through nonparametric regression. This new model has shown some

encouraging results on identifying the sources (Pazokifard, 2007; Henry, 2008). NTA gives a map of the average concentration at the receptor when the air passes over each point on the map. This NTA map shows the relatively high concentration areas as possible pollution sources. However, the identified areas do not really indicate the real sources since the trajectories may over bending and over expending the real source. Figure 1 shows that since trajectories are bending over, it may cause the real source obscure. The main purpose of this study is to refine and to improve the NTA method to identify and quantify the sources of pollution.

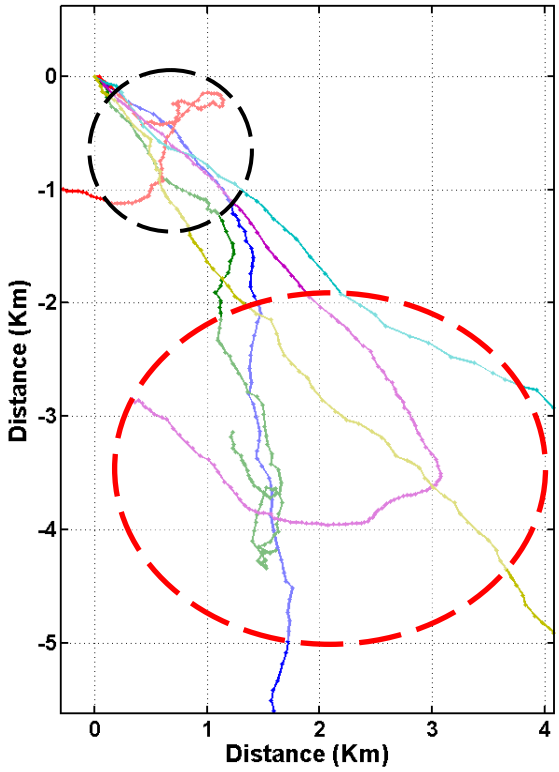


Figure 1– A typical trajectories curve. The black dash circle points out a real source of pollution. The Red dash circle is the artifacts source of pollution.

C. Proposed Methodology

One way to extend the NTA method is the Point Source Response (PSR) method. The

way to make PSR is to associate the real pollution concentration data with the designed point sources by multiplying a Gaussian decay factor, Q_k the suffix k is the kth point source, and this will go into details in chapter 4 data analysis method.

The basic assumption of PSR is the NTA map of real sources can be linearly combined by point source responses from numbers of point sources. In addition, obviously the influence of the real source is negative related to the distance of test point sources. The closer the trajectories to the point source will have higher impact by the point source. In Figure 2, it shows that red spot is an actual source, and points A to D are test point sources. The point source response of A will be the strongest impacted by real source, points B and C might be in the same influence, and point D will be the slightest impacted by the real source.

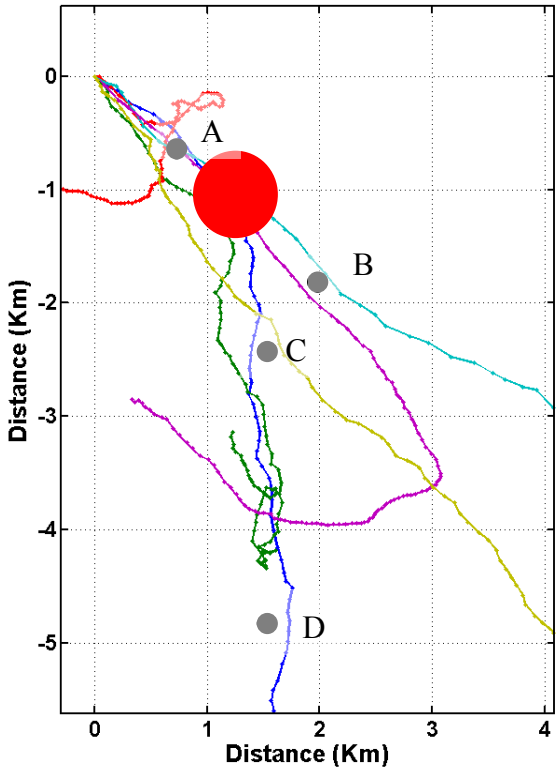


Figure 2– The red spot is a real source, points A to D are test point sources.

Applying NTA on each point source, the NTA map for point source at each point on a grid is calculated. For the next step is to choose a best fits of weighting coefficients by Principal Components Regression (PCR) for each PSR map. A similar NTA map for real data can be produced by a weighted sum of these PSR maps. Regarding how to apply PCR to determine the weighting coefficients will be mentioned in Chapter 4.

Physically, these weighting coefficients are related to the contribution of each point source. By further analyzing these weighting coefficients, a sector apportionment for a specific angle from the receptor can be determined.

D. Content of Each Chapter

The content to this research is presented in the following order.

Chapter 2 covers a discussion of the previous research and background information on nonparametric regression analysis method and also trajectory techniques in air quality research.

Chapter 3 is a description of data. This chapter includes details on monitoring sites, their locations and parameters used in this research.

Chapter 4 is the detail on the mathematics behind the hybrid model. It covers a description and general overview of the method. It also gets into mathematics details of nonparametric regression analysis, smoothing, error analysis, calculating back trajectories, and detailed methodology of point source response techniques, and data screening and how to choose data among many available options.

Chapter 5 is the presentation of results of applying the Point Source Response technique to identify and quantify the local sources of some specific angles from the monitoring site. The results presented for two regions, Long Beach with pollutant SO_2, and Rubidoux with emphasize on PM_{10}.

Chapter 6 covers conclusions based on the results of research and suggestions to the future developments of this model.

Chapter 2

Review of Literatures

A. Nonparametric Regression analysis method for air quality data

Nonparametric regression analysis has been developed and used for decades. This statistical method has been widely applied on several different fields such as biology, agriculture, social science, and engineering, but it has not really engaged in air quality or atmospheric studies.

Henry et al. (2002) has first shown the effectiveness of nonparametric regression analysis as the technique in locating emissions sources when applied on wind direction

data and pollutant concentrations. Further, Henry's method was expanded to include both wind direction and speed in simulating pollutant concentrations involving major international airports (Yu et al. 2004). This study showed that the method could be used to identify the influence of airports on local air quality. When each of the SO_2, CO, and NOx measurements were examined closely, the overall air quality upon contributions of emissions from aircrafts and from the ground equipment or vehicle could be identified separately.

B. Trajectory Techniques in Air Quality

Applications of trajectory techniques are extensively appeared in meteorology, climatology, environmental science, and law enforcement. Back trajectory method indicates back tracking the past path of pollutants to determine and locate the sources of emissions. Associating with the results of this technique and air quality data

gathered in monitoring stations, it helps us to estimate and locate the local pollution hot spots and clean air zones.

There are numerous researches and publications on back-trajectory methods. For examples Subhash el al., measured the concentration of polychlorinated biphenyl (PCB) at a site on the Lake Superior shoreline and analyzed in conjunction with back-trajectory to assess the contribution of long range transport of measured PCB concentrations. The results indicated rapid transport from urban and industrial regions well south of sampling site (Subhash el al, 1999).

Sturman et al. applied atmospheric modeling to delimitation of clean air zones for urban areas in New Zealand. The results indicate in spite of low wind speed, the air tend to travel from a significant distance outside the boundary of the city. It shows that cold air typically travels up to 20 km from the west into the city during air pollution events. (Sturman et al, 2002)

Draxler et al. in 1997 developed a hybrid method between Eulerian and Lagrangian approaches which is called Hybrid Single- Particle Lagrange Integrated Trajectory (HYSPLIT). This method is designed for quick response to atmospheric emergencies, diagnostic case studies, or climatological analysis. (Draxler et al., 1998). Escudero et al., applied this method to determine the contribution of northern Africa dust source areas to PM_{10} concentrations over the central Iberian Peninsula. It correctly showed the concentrations profiles and direction of sources of PM_{10} in this area. (Escudero et al., 2006)

C. Source Apportionment Methods in Air Quality

In order to determine and quantify the relative contributions of various source types to ambient air pollutant concentrations to a location of interest, researchers applied and developed many different air quality models to achieve this purpose.

Generally, air quality models can be classification in two categories: source-oriented air quality models, receptor-oriented models, and new hybrid source- receptor models.

Source-oriented models, also called traditional dispersion models, include Lagrange, Euler and Gaussian plume models. To have accurate results, they need detailed information on emission sources, local meteorological conditions and detailed chemical reactions. The information on the emission sources includes emission rate, pollutant compositions, and source locations.

Receptor-oriented models are such as chemical mass balance (CMB), multivariate receptor models, positive matrix factorization (PMF), backward trajectory models, potential source contribution function (PSCF) and UNMIX. These models use chemical compositions of pollutants as fingerprints to locate the emission sources and apportion the contributions of various air pollution sources.

Most of the above air quality models can be used for source apportionments. These methods including investigation of the spatial and temporal characteristics of data; cluster, factor, and other multivariate statistical techniques; positive matrix factorization (PMF); UNMIX; the chemical mass balance (CMB) model; and trajectory analysis tools such as Gaussian Trajectory transfer coefficient model (GTx); Trajectory Mass Balance (TrMB); potential source contribution functions (PSCFs); Hybrid Single- Particle Lagrange Integrated Trajectory (HYSPLIT).

Kim et al. analyzed and improved source identification in daily integrated $PM_{2.5}$ composition data including various individual carbon fractions collected in Atlanta area by positive matrix factorization (PMF) method. He indicated that the temperature resolved fractional carbon data can be utilized to enhance source apportionment study, especially with respect to the separation of diesel emissions from gasoline vehicle sources. Conditional probability functions using surface wind data and identified source contributions aid the identifications of local point sources. (Kim et al., 2004)

Srivastava et al. demonstrated Chemical Mass Balance (CMB) model to source apportionment of ambient VOCs in Delhi City. His research indicated that emissions from diesel internal combustion engines dominate the air quality in Delhi. Vehicular exhausted and evaporative emissions also contribute significantly to VOCs in ambient air. (Srivastava et al., 2004).

Lewis and Henry et al. showed another receptor model, UNMIX model to estimate source apportionment to analyze a 3-year $PM_{2.5}$ ambient aerosol data set collected in Phoenix, AZ. The analysis generated source profiled and overall average percentage source contribution estimates for five source categories: gasoline engines (33±4%), diesel engines (16±2%), secondary SO42- (19±2%), crustal and soil (22±2%), and vegetative burning (10±2%). (Lewis and Henry et al, 2003)

Tsai et al. used chemical characteristics of winter aerosol at four sites in southern Taiwan, and the Gaussian Trajectory transfer coefficient model (GTx) to identify the major air pollutant sources affecting the study sites located in southern Taiwan. The

most important constituents of the particulate matter (PM) by mass were SO_4^{2-}, organic carbon (OC), NO_3^-, elemental carbon (EC) and NH_4^+, with SO_4^{2-}, NO_3^-, and NH_4^+ together constituting 86.0–87.9% of the total $PM_{2.5}$ soluble inorganic salts and 68.9–78.3% of the total $PM_{2.5–10}$ soluble inorganic salts, showing that secondary photochemical solution components such as these were the major contributors to the aerosol water-soluble ions. (Tsai el al., 2006)

Gebhart et al. used Trajectory Mass Balance (TrMB) which is basically a receptor model to estimate source apportionment of particulate sulfur measured at Big Bend National Park, TX. The results quantify the apportionment of the mean sulfate source from different regions of United States (7-26% from the eastern US, 12-45% from Texas, and 3-25% from the western US) and Mexico (39-50%) (Gebhart et al, 2006).

Lucey et al., studied the source receptor relationships by potential source contribution functions (PSCFs) for fourteen chemical species found in precipitation collected at Lewes, Delaware. This results indentified the likely emission sources for these

chemical species includes oil- and coal power plants, incinerators, motor vehicles, and iron and steel mills all over the areas of the Eastern United states.(Lucey et al., 2001)

D. Uncertainties Estimates

As a new air quality model is developed, it is necessary to do uncertainty estimate and give the confidence intervals. The resulting estimates of pollution source profiles have error and frequently the uncertainties are obtained under an assumption of independence. In general, traditional methods such like Bootstrap and Jackknife approaches are widely accepted. They provide confidence intervals and standard errors for receptor model profile estimates under some assumptions of independence. For example, Henry used Bootstrap resampling methods to determine the uncertainties for UNMIX air quality models. (Henry et al., 1999; Henry, 2000)

Spiegelman et al., compared both Bootstrap and Jackknife methods on a receptor model, constrained nonlinear least square (CNLS) method. He suggested that the

jackknife approach tends to produce larger standard error estimates and wider confidence intervals than the Bootstrap method done under the assumption of independence. (Spiegelman et al., 2007)

In addition there is an application of another more intuitive way, such like Henry et al. and Yu et al., calculated confidence intervals for 2 and 3 dimensions nonparametric regression from formulae based on the asymptotic normal distribution of kernel estimates. (Henry et al., 2002; Yu et al., 2004)

Chapter 3

Data Description

A. Monitoring Site

In this study, the following two sites were chosen for local pollution sources study: North long Beach, and Rubidoux. These monitoring sites were chosen based on their geographical and meteorological conditions in covering these various setting and availability of data on necessary air quality and meteorological parameters.

The data used in the objective for the preliminary works are done by the South Coast District Air Quality Management District (SCAQMD). These data are 1 minute concentrations along with wind direction and wind speed. These monitoring sites were

chosen because of particularly conditions of local air quality and meteorological parameters. North Long Beach is specifically important because of significant emission sources around it, such as the port of Long Beach, the port of Los Angeles, Long Beach airport, major freeways and intersections, and refineries in this area. Because these emission sources produce relative high amount of SO_2, this study will focus on analyzing these two pollutants. Rubidoux on the other hand is important due to the fact that there have always been air quality standard violations in this area. In addition, since new constructions are continuously built, particulate matters are far from ignoring in this area. Hence, for Rubidoux, TEOM[1] PM_{10} will be specially discussed. The locations of cities and monitoring stations are shown in Figure 3 to Figure 5. More detailed information of these monitoring sites is brought in the following Table 1.

---

[1]TEOM : Tapered Element Oscillating Microbalance.

Figure 3 – Map of the monitoring site locations: A = North Long Beach, B = Rubidoux

(Source of map: Google Earth)

Monitoring Site	UTM[1] Latitude	UTM Longitude
Los Angeles International Air Port (LAX)	386804.14	3782435.57
West Los Angeles	365553.74	3768718.45
Lynwood	388084.93	3754931.98
*North Long Beach	390004.94	3743232.27
South Long Beach	391198.17	3739738.62
Crestline	474624.90	3788954.84
Pomona	430792.53	3769801.13
*Rubidoux	461550.30	3761901.09
San Bernardino	474723.00	3774018.33
Upland	441968.01	3773853.57

Table 1 – The Monitoring sites UTM coordinates.

The* means the sites chosen for case study in this research.

[1]UTM : Universal Transverse Mercator grid system. All of the sites are in zone 11 S.

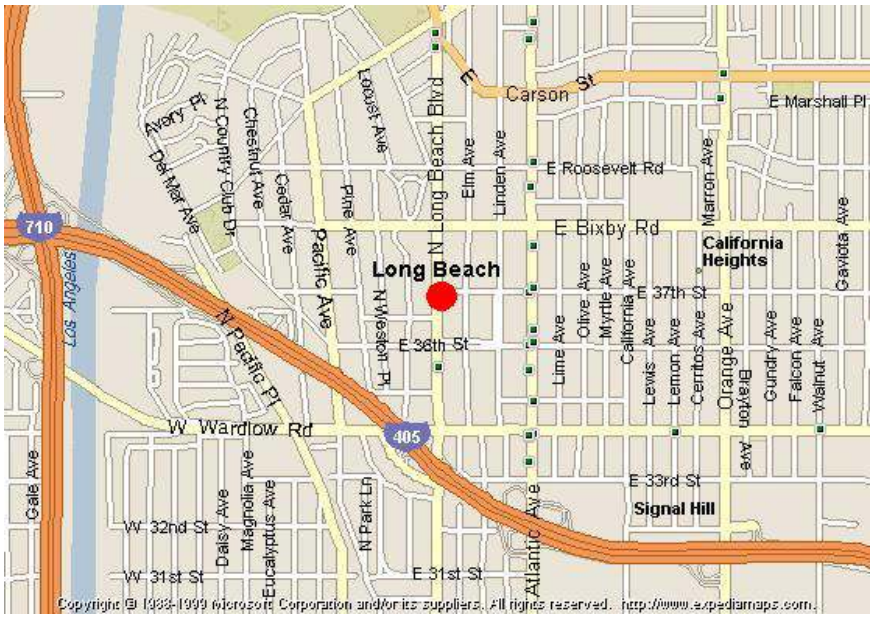


Figure 4 – A detailed view of North Long Beach Site Location; Address: 3684 Long Beach Blvd. Long Beach, CA90807. (Source of map: California Air Resource Board, CHAPIS webpage)

The city of Long Beach is located approximately 20 miles south of downtown Los Angeles. A monitoring site was placed in a busy area near the I-710 and I-405 freeways. It is located 0.5 miles north of I-405 and about one mile east of I-710 (Refer to Figure 1 and 2). There are several important sources of emission in this area. Large oil refineries, petroleum product chemical plants, and power plants are located further away to the west and southwest direction from the monitoring station. Considering their size, they can be one of the major contributions to the emission sources. Regarding the harbor area, it is considered the busiest port in U.S. Therefore, the

harbor area is also an important contributor to the pollution of the area. Various types of ships such as cargo transportation and cruise travel ships initiate or end their trip from or to this port. The port also brings high traffic of heavy-duty vehicles for transportation of the containers. In addition, the port needs heavy equipment and facilities for loading and unloading the cargo and passengers or generally to run the port. Besides, another emission sources would be Long Beach airport 2 miles southeast of the site.

The Long Beach displays typical coastal weather patterns. A typical diurnal pattern of wind speed and direction results from a sea breeze during the day time, which blows from a western direction and from a north to northeastern direction during the night time, which is also called “drainage flow”. The direction of wind changes during the dawn and dusk, which results in reduction of wind speed. This diurnal wind pattern remains relatively similar throughout the year except for the fall. During the fall, the southern California region experiences a strong wind pattern called Santa Ana wind, which strongly blows from north to northeastern pretty much throughout the whole day.

During the Santa Ana wind period, because of the high velocity of air the air quality of the southern California increases significantly, which also leads into better visibility. However, this type of condition occurs only during a specific season and only for a relatively short period.

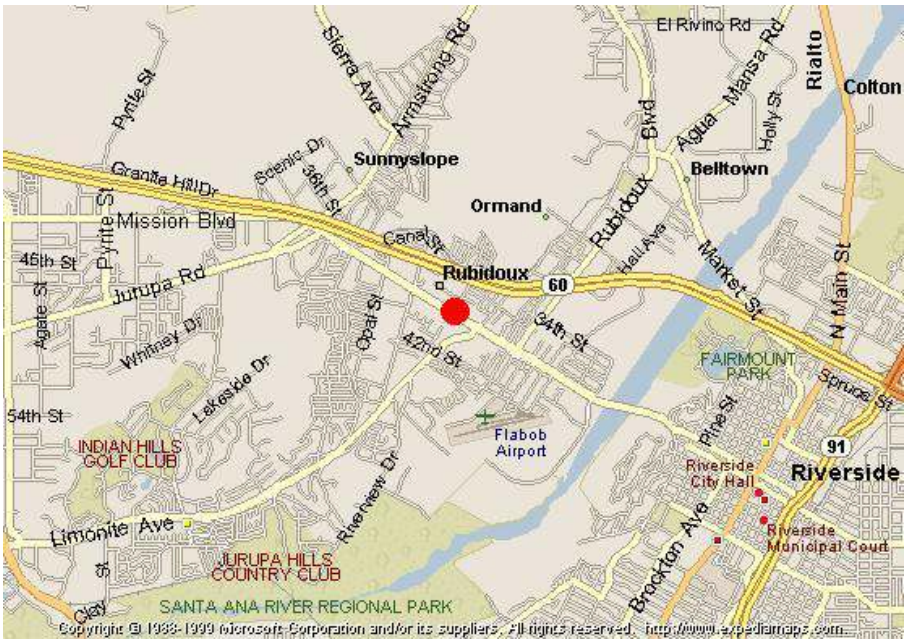


Figure 5 – A detailed view of Rubidoux Site Location; Address: 5888 Mission Blvd. Riverside, CA92509(Source of map: California Air Resource Board, CHAPIS webpage)

Riverside County is located within three air basins: South Coast Air Basin (SOCAB), Salton Sea Air Basin (SSAB), and the Mojave Desert Air Basin (MDAB). U.S.

government established 11 monitoring stations located throughout these three air basins. The data chosen for this research are from the Rubidoux in the South Coast Air Basin.

Rubidoux is located about 50 miles east of Downtown Los Angeles. The monitoring site is located about 0.5 miles south of CA/SR-60 (Pomona Freeway, refer to Figure 1 and 3). The interested emissions sources of the area are nearby the Pomona freeway. Prevailing wind in the Riverside area is from the west and southwest. These winds are due to the proximity topographies, such as the coastal, central regions, and the presence of Sierra Nevada Mountains which is a natural barrier to the north. The dominant daily wind pattern is a daytime sea breeze (onshore breeze) and a nighttime land breeze (offshore breeze). They are broken only occasionally by winter storms and infrequent strong Santa Ana winds from the Great Basin, Mojave, and deserts to the north. The “flushing” phenomenon tends to be more during the spring and early summer. In these months of the whole year, the basing can be flushed of the pollutants by ocean wind during afternoon. However, during the late summer and winter months, the flushing is less noticeable because of lower wind speeds and the earlier occurrence

of offshore wind. As a result, pollutants are trapped and begin to accumulate during the night and the following morning.

B. Parameters used in the study

Among the numerous parameters that were constantly measured at these sites, the following pollutants and meteorological parameters were used for this study including their units of measurement: SO_2 in ppb, wind speed in miles per hour, and wind direction in azimuth degree from north Long Beach site and TOEM PM_{10} in $\mu m/m^3$, wind speed and wind direction from Riverside site.

SO_2 is used from the Long Beach data because of many interested SO_2 emissions sources are considered as a dominant pollution. PM_{10} has been chosen as the main pollutant in Riverside because its concentration exceeds federal standards and has the highest value in SOCAB.

C. Data Collection and Distribution

The air quality data needed for this study was obtained from the South Coast Air Quality Management District (SCAQMD). SCAQMD is a regional government agency that oversees air quality monitoring and necessary regulatory actions to improve air quality in the Los Angles air basin. Numerous routine monitoring stations are maintained by the SCAQMD and CARB for collecting are quality data throughout the region.

Continuous measurements of air quality and meteorological data were made at these monitoring sites. At every minute, these measurements were recorded by an instrument and transferred to headquarters of the SCAQMD. This database supplies U.S. EPA AirData (formerly known as Aerometric Information Retrieval System, AIRS) and the California Air Resources Board (CARB) in which the air quality database is part of a network where the public can visit and access portions of the same data used in this research.

Chapter 4

Data Analysis method

In this chapter, a new source apportionment method will be described. The method mentioned here is an extension of the application of nonparametric regression in air quality modeling (Henry et al., 2002; Yu et al., 2004). This model starts from calculating back trajectories using the wind speed and direction data from one or more sites. Once the trajectories have been calculated, the concentration of pollutant at the time of arrival at the monitor is associated with each point on the trajectory. Thus, an expected value of concentration associated with back trajectories can be calculated. By further applying the “point sources response” for each point at a known grid, the estimated source apportionments can also be found out.

A. Nonparametric Regression Methodology

a. Nonparametric Regression

Nonparametric regression is a method of estimating the mean value of a dependent variable given the value of one or more predictor variables (Härdle, 1990; Wand and Jones, 1995). Ordinary regression does the same thing, except that a functional relationship between the dependent and predictor variable is assumed to be known. The most common functional forms are linear, polynomial, exponential, and logarithmic. The data are then used to estimate the unknown parameters of the function; for example, the intercept and slope of a linear function. Thus, ordinary regression can be called parametric regression. However, there is an obvious drawback to the parametric regression. Since it needs to be considered which functional form is the most appropriate one, at least approximately, there is a danger of reaching incorrect conclusions in the wrong assumption of regression analysis. The rigidity of parametric

regression can be overcome by using nonparametric regression because it does not need to consider which the fittest functional form is. Nonparametric regression is a method of estimating the relationship between the dependent and predictor variables without making any assumptions about the functional form of the relationship or the statistical distribution of the data. The motivation for a nonparametric approach is straightforward and intuitive. This is sometimes referred to as “letting the data speak for themselves”. It is especially true when facing a large amount of data, such as air quality data. For example, the Figure 6 shows the SO_2 concentration measured at North Long Beach site during entire January 2006. It is very difficult to apply parametric analysis, such as linear regression to find the trend of concentrations versus wind direction since the data distribution is not only too concentrated at lower concentrations but also not easy to figure out the trend of this figure.

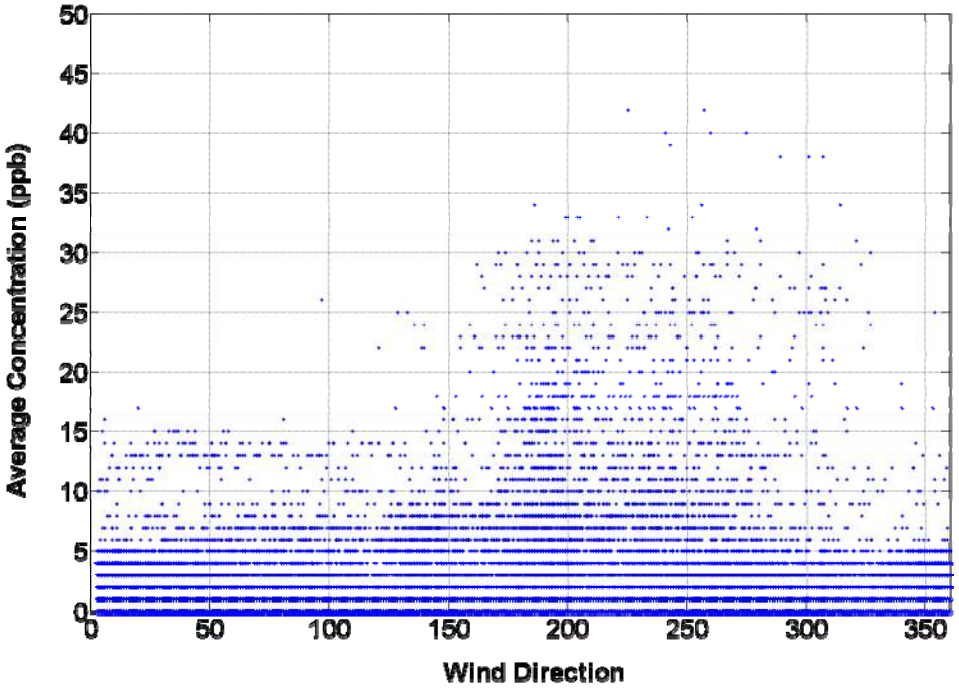


Figure 6 – SO_2 observed concentration vs. wind direction at North Long Beach site during January 2006.

Using nonparametric regression would be possible to deal with this issue. Histogram graph is the oldest and the most widely used nonparametric analysis shown as figure 7, which is with a bin width equals 10-degree azimuth. It is a step function with heights being the proportion of average value contained in that bin divided by the width of the bin. The choice of the bin width is usually called smoothing parameter since it controls how smooth the histogram chart looked like while a larger bin width results in a

smoother looking histogram. This preliminary nonparametric analysis shows at least a distinguishable trend although it does not really give us a very satisfied result here.

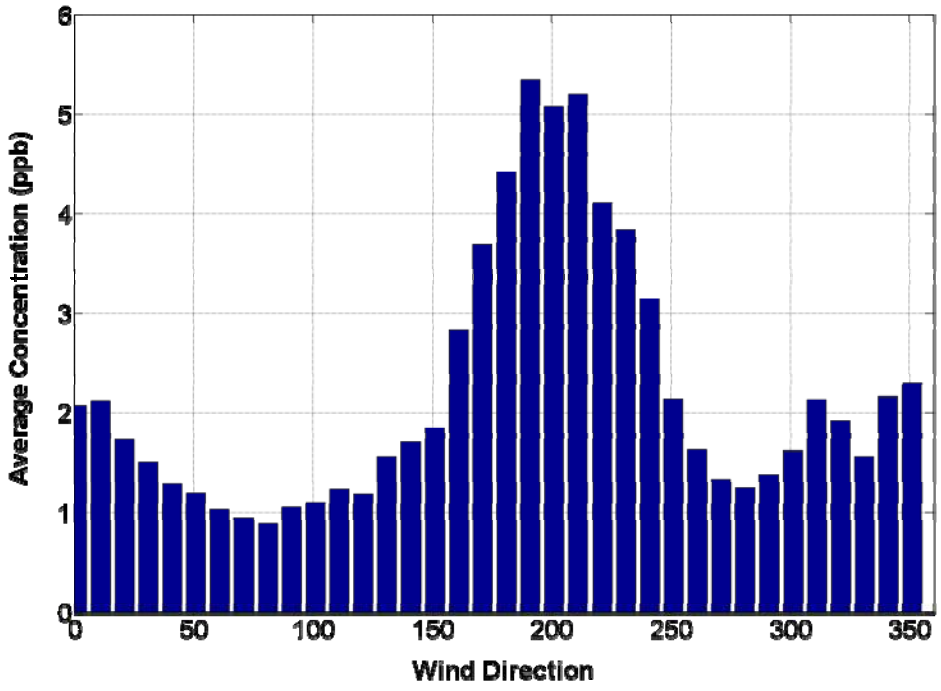


Figure 7 – Histogram with bin width 10 degree azimuth on SO_2 observed concentration vs. wind direction at North Long Beach site during January 2006.

Considering another nonparametric smoothing factors which is called kernel density estimator, the result obviously improves some of the problems of a simple histogram. The results are shown in figure 8 and figure9. Figures are presented by nonparametric regression with Gaussian kernel smoothing. In figure 9, the centerline is the smoothed fit on the observed data and the dashed lines above and below the centerline show a 95% confidence interval. The detailed mathematical introduction of kernel smoothing will be discussed in the following section. As we see, the smooth line is easy to follow and the data are also easy to understand. It does give us a better choice to present the trend of pollutant concentration versus wind direction.

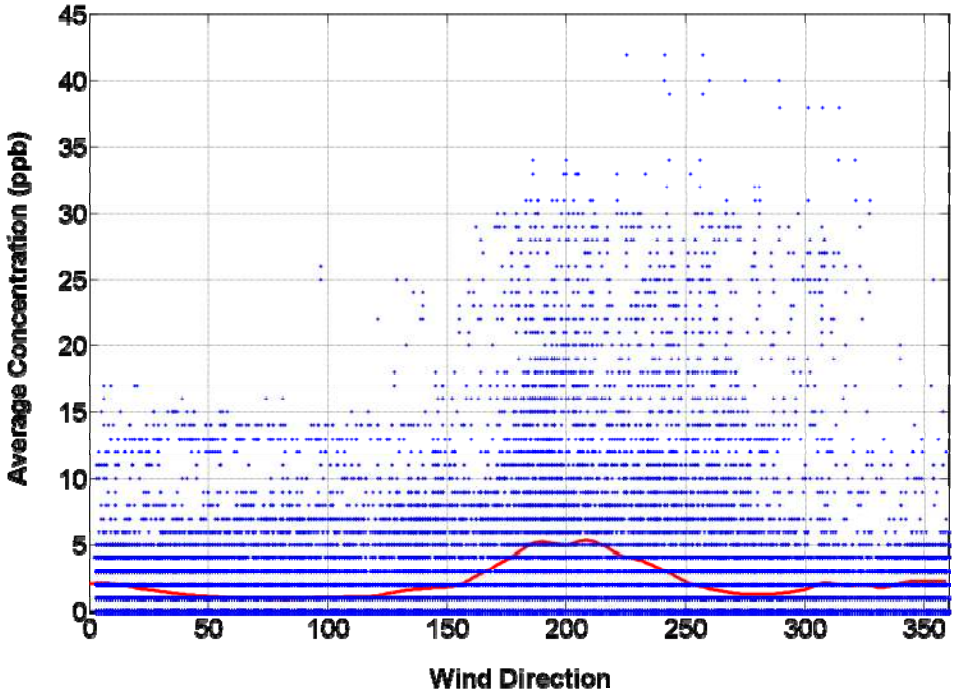


Figure 8 – Nonparametric regression on SO_2 observed concentration vs. wind direction at North Long Beach site during January 2006.

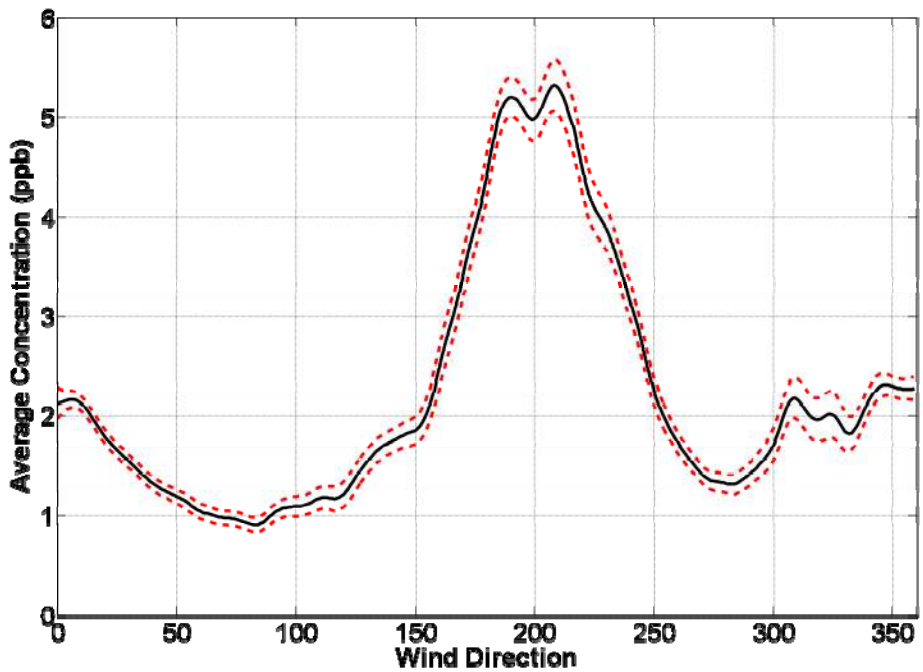


Figure 9 –Nonparametric regression on SO_2 observed concentration vs. wind direction at North Long Beach site during January 2006. The dash lines are 95 % confidence intervals.

Using nonparametric regression to analyze the data, we will have four advantages (Härdle, 1990). First, it offers a generally more flexible method in exploring a general relationship between variables. Second, it offers the observed data speaking by themselves without making assumptions of rigid parametric method. Only using smoothing methods enable to identify the general trend of complex data, which is barely analyzed by simple parametric methods. Third, by discussing the bias and

variance of the data set, we can figure out the distribution of effect main body of data. Fourth, this method is more flexible in substituting missing data or interpolating data from adjacent data points in the main trend rather than using the whole data set.

b. Kernel Smoothing

The relationship between pollutant concentrations and wind direction that is given in the Figure 9 can be calculated by using the nonparametric kernel regression estimator given in the below equation.

$$E(C \mid \theta, \sigma) = \frac{\sum_{i=i}^{N} K\left(\frac{\theta - W_i}{\sigma}\right) C_i}{\sum_{i=i}^{N} K\left(\frac{\theta - W_i}{\sigma}\right)} \qquad \text{(a)}$$

Where, E $(C|\theta,\sigma)$ is the expectative calculated average concentration, which is a weighted average from the concentration of a pollutant for a particular wind direction. C_i is the observed average concentration for the period starting at t_i where i =1… n

observations. In addition, W_i is the resultant wind direction for the ith time period. θ is the wind direction, σ represents smoothing parameters, usually called the bandwidth or window width. K is the kernel density estimator, which is a function satisfying $\int_{-\infty}^{\infty} K(x)\,dx = 1$,or called the kernel weight since it can be looked as weighting factor in this function (Wand and Jones, 1995). The above equation is usually called Nadaraya-Watson estimator (Härdle, 1990). This is known to be consistent, that is, as sample size increases that vale of estimate will converge to the true value (Henry et al, 2002).

Applying in the same concept, we expand equation (a) to 3-dimentional nonparametric analysis as shown by equation (b).

$$E(C \mid \theta, u) = \frac{\sum_{i=1}^{N} K_1\left(\frac{(\theta - W_i)}{\sigma}\right) K_2\left(\frac{(u - U_i)}{h}\right) C_i}{\sum_{i=1}^{N} K_1\left(\frac{(\theta - W_i)}{\sigma}\right) K_2\left(\frac{(u - U_i)}{h}\right)} \qquad \text{(b)}$$

Where, U_i is the resultant wind speed, h represents smoothing parameters for K_2, and

the rest of the notations remain the same as (a). This equation results in E (C| θ, u), which is the observed average concentration of a pollutant associated with a particular wind speed and direction pair. In the equation above, K_1 and K_2 are the Gaussian kernel estimator and the Epannechnikov kernel estimator respectively. The functions are shown following.

The Gaussian kernel estimator:

$$K(x) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{x^2}{2}) \qquad -\infty < X < \infty,$$

The Epannechnikov kernel:

$$K(x) = 0.75 \times (1 - x^2) \qquad -1 < X < 1,$$

Both of these kernels will give higher weight to observations near x and less weight to observations further away. There are many possible choices for kernel density estimator K, the major difference between the two which we are using here is that the

Gaussian kernel is defined over an infinite domain and the Epannechnikov kernel is defined over a finite range. The Gaussian kernel is preferred for wind direction since it is defined over an unbounded range, and the Epannechnikov kernel is preferred for data limited to a finite range like wind speed. An important decision for using nonparametric regression method is to choose appropriate smoothing parameters σ and h, or, equivalently defined as the FWHM (Full Width at Half Maximum). If the FWHM is too large, the curve will be too smooth and peaks could be lost or indefinable. On the contrary, if chosen FWHM is too small, the curve will lead to a relatively jagged figure, too many meaningless peaks dominated by noise, or large peaks resolve into false multiple peaks.

FWHM is an intuitive measurement of the window width for the kernel estimators. It is simply the full width of the peak in K measured at the point where the curve has fallen to half of its value at the peak. For the Gaussian kernel, FWHM and the smoothing parameter σ are related by:

$$\sigma = \frac{\text{FWHM}}{2\sqrt{-2\ln(0.5)}}$$

For the Epannechnikov kernel, the smoothing parameter h is given by:

$$h = \frac{\text{FWHM}}{\sqrt{2}}$$

Methods to estimates value of the smoothing parameters that are optimal under certain assumption are available (Wand and Jones, 1995). There are also other statistical methods mentioned in the book by Härdle (1990). Throughout this study, values for FWHM were based on the experience practically choosing the optimal values. The FWHM values represented in this study are chosen to be 15 degree for the Gaussian kernel and 3 mph for the Epannechnikov.

The variance of the average concentration in equation (b) is estimated by

$$\mathrm{Var}(E(C\mid\theta,u)) = \|K_1\|^2\|K_2\|^2 \frac{\sum_{k=1}^{N} K_1\left(\frac{(\theta - W_k)}{h_1}\right) K_2\left(\frac{(u - U_k)}{h_2}\right)(C_k - E(C\mid\theta,u))^2}{\left[\sum_{k=1}^{N} K_1\left(\frac{(\theta - W_k)}{h_1}\right) K_2\left(\frac{(u - U_k)}{h_2}\right)\right]^2} \qquad \text{(c)}$$

where,

$\|K_1\|^2 = \int_{-\infty}^{\infty} K_1^{\,2}(x)dx = 1/2\sqrt{\pi}$, for the Gaussian kernel, and

$\|K_2\|^2 = \int_{-\infty}^{\infty} K_2^{\,2}(x)dx = 0.6$, for the Epannechnikov kernel.

The above result is derived from general expressions given in Wand and Jones (1995).

Figure 10 is the typical 3-D Nonparametric Regression application graph to illustrate SO_2 concentration in ppb v.s. wind speed and wind direction for North Long Beach during July 2005

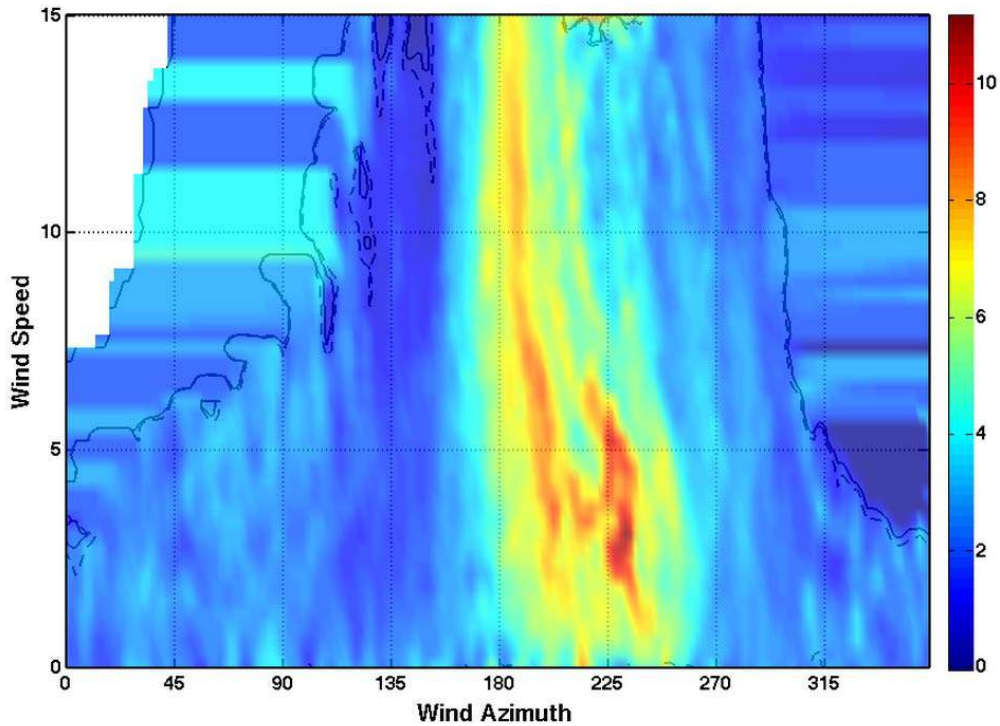


Figure 10 –A typical 3–D Nonparametric Regression graph illustrates SO_2 concentration in ppb v.s. wind speed and wind direction for North Long Beach during July 2005 fwhm = [5 1]

Sometimes it is more convenient to work with Cartesian coordinates instead of polar coordinates. If the positive x-axis is east, then the wind speed and direction are replaced with x, y coordinates by

$$\begin{aligned} x &= u\cos\phi, \\ y &= u\sin\phi, \\ \phi &= \pi \,\mathrm{mod}(450-\theta,360)/180 \end{aligned} \tag{d}$$

Where the last equation converts the azimuth angle in degrees clockwise from north to the mathematical angle ϕ in radians counterclockwise from +x-axis; here mod(a,b) is a modulo b. in this case, since both predictors are bounded, K_2 in the above equations is taken to be the Epanechnikov kernel rather than the Gaussian kernel, otherwise the equations are be same as above with θ and u replaced by x and y.

B. Nonparametric Back Trajectory Analysis

a. Using Single Monitoring Site of meteorological data

The following is the description of the trajectory models to be used in this research using data from one monitoring site only. As shown in Figure 11, wind azimuth is the direction of wind coming from, which is measured clockwise from north. To calculate the x (east west) and y (north south) coordinates of the wind direction; the azimuth angle must be converted to the usual mathematical definition of angle which is

measured counterclockwise from the x-axis. If the azimuth angle is Z and the mathematical angle is θ, then:

$$\theta = \text{mod}\ (Z - 450, 360), \text{ and}$$

$$Z = \text{mod}\ (\theta - 450, 360)$$

As defined above, θ is between 0° and 360°.

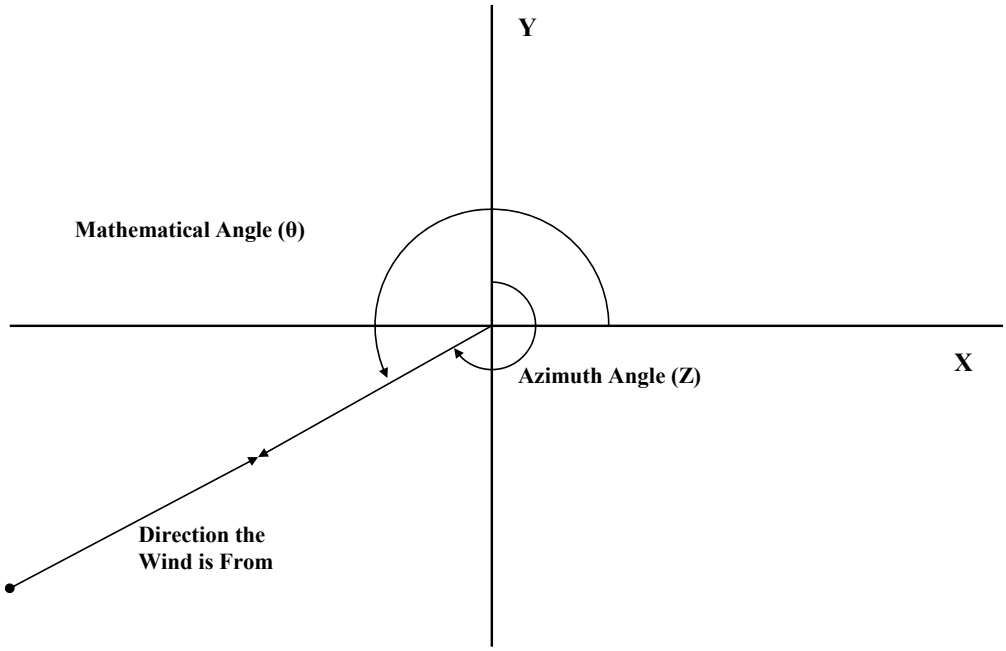


Figure 11 – Definition of angles used for trajectory calculations.

If the wind speed is u, the x and y coordinates of the wind velocity at time t_k are then:

$$v_x(t_k) = u(t_k)\cos(\theta(t_k))$$

$$v_y(t_k) = u(t_k)\sin(\theta(t_k))$$

Then the x and y coordinates of the points on the back trajectory starting at time t_j are

$$x_k(t_j) = \sum_{i=0}^{k} v_x(t_{j-i})\Delta t$$

$$y_k(t_j) = \sum_{i=0}^{k} v_j(t_{j-i})\Delta t$$

$$k = 1,\ldots, N$$

Where Δt is the time step, that is the time between measurements and N is how many steps backward in time are taken. More complex schemes to calculate back trajectories using wind and other meteorological data from additional site can also be used. Each point on the trajectory is associated with c_j the concentration at time t_j when the air arrives at the receptor. Finally, all the points from the set of all the trajectories of interest starting at all possible times along with the associated concentrations are assembled in a set of ordered triples (x_i, y_i, c_i), where the suffix i ranges over all the points of all the trajectories (Henry, 2008). Figure 12 is a typical back trajectory result made from meteorological data of single monitoring site at North Long Beach.

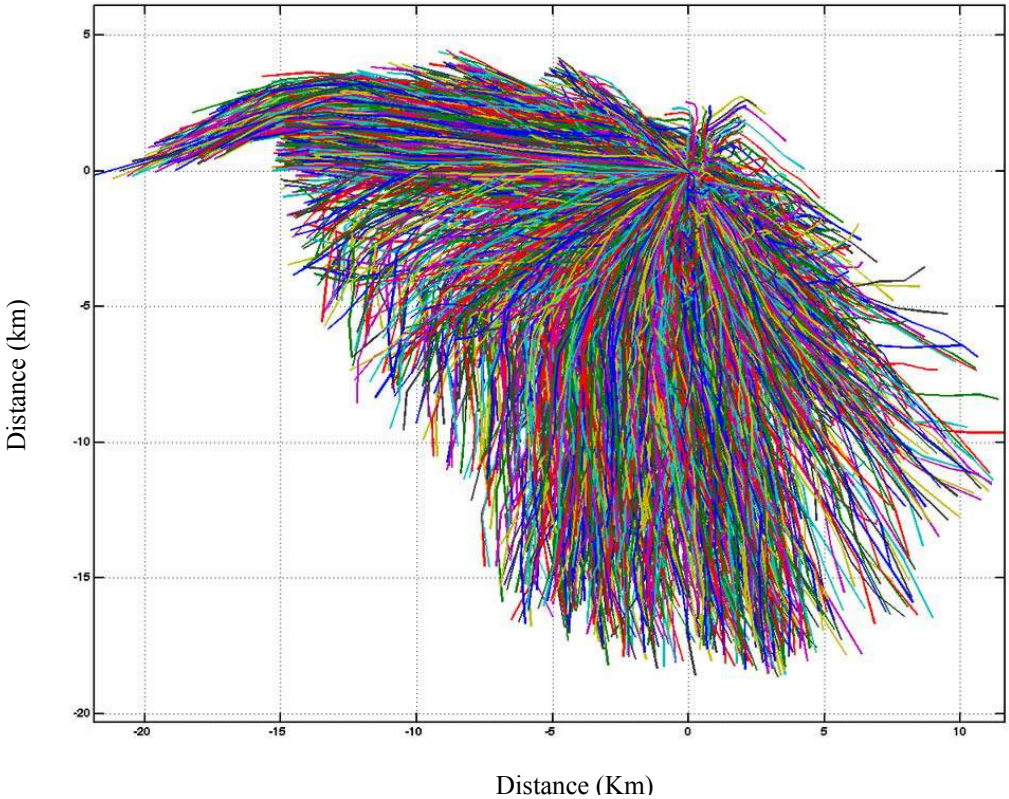


Figure 12 – A typical 2– hr back trajectories result from single site of WS and WD during July 2005 at North Long Beach site.

b. Using Multi Monitoring Sites of meteorological data

The further from the monitor, the more unreliable the trajectories become. In order to overcome this problem, gathering more data from multi monitoring sites is a worth way to consider. Meteorological data from several sites are combined with each other

and the wind vectors are calculated by calculating the weighted average of horizontal and vertical vectors of wind speed and direction from different monitors with respect to the distance from them. X (east-west vector) and Y (north-south vector) are calculated as follows:

$$X(t_k) = \frac{\sum\left(\frac{1}{r_j^2} \times x_j(t_k)\right)}{\sum\left(\frac{1}{r_j^2}\right)}$$

$$Y(t_k) = \frac{\sum\left(\frac{1}{r_j^2} \times y_j(t_k)\right)}{\sum\left(\frac{1}{r_j^2}\right)} \qquad \text{(e)}$$

By this calculation, X and Y get better weight in averaging when they are closer to any of the monitors.

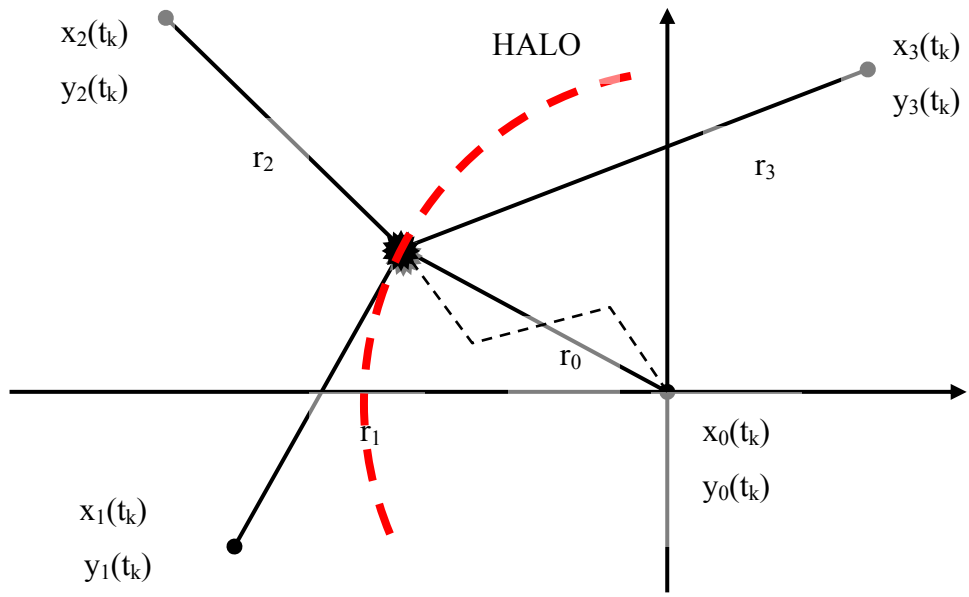


Figure 13 – Schematic of weighted average for using multiple monitoring sites.
The Red dash line means HALO

In using multi monitoring sites analysis, there are some basic assumptions should be noted. On one hand, when the air parcels are close enough to a monitoring site, it makes sense to ignore the effect of the wind speed and wind direction data from other monitoring sites. In such a case, only meteorological data from the closest monitoring site will be used to calculate the trajectories. While calculating the back trajectories using multi sites, a radius of one kilometer were employed, which is called “HALO” shown in Figure 13 with red dash line. It means that when the air parcel is in one kilometer of radius of any monitoring sites, the effect of other sites out of the range of

radius are neglected and not brought into calculations. On the other hand, if the monitoring site is too far from the air parcel, it does not make sense to affect the path while calculating the back trajectories. A reasonable range, 50 kilometers, was used in this research to calculate the trajectories. These values were defined based on experience (Pazokifard, 2007). Figure 14 is a typical 2 hr back trajectories results from meteorological data of multiple sites during July 2005 at North Long Beach site with HALO 1500 m.

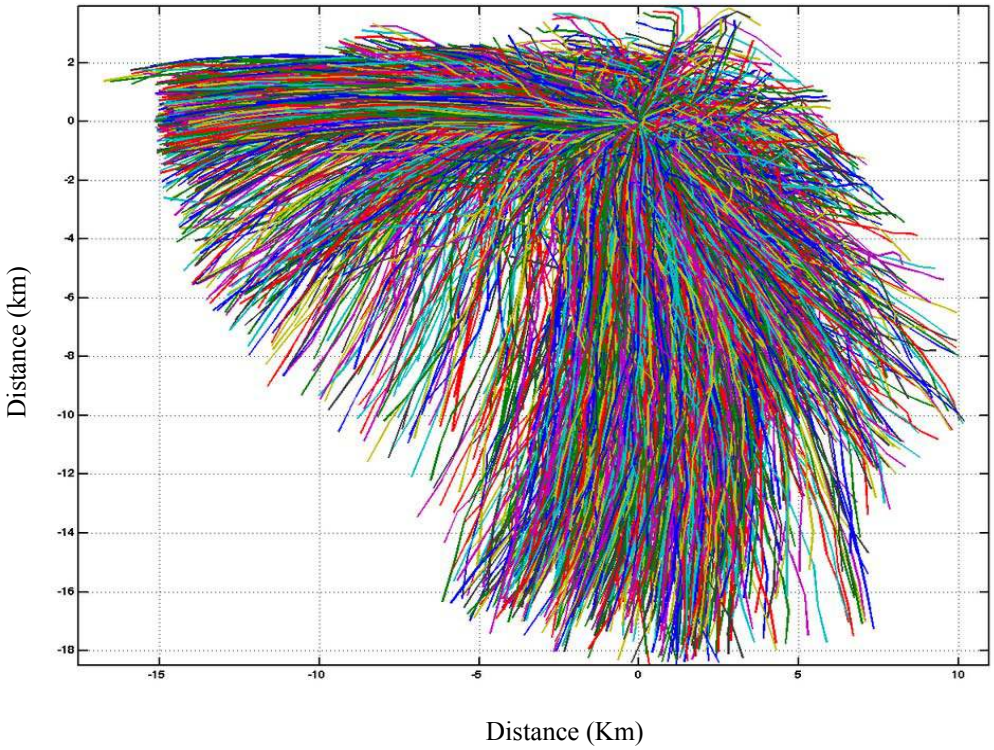


Figure 14 – A typical 2– hr back trajectories result from meteorological data of multiple sites during July 2005 at North Long Beach site with HALO = 1500 m.

c. Smoothing analysis

We apply kernel smoothing which is mentioned in the previous section to estimate the average concentration at the monitor if the air passes over a point on the map from the set of trajectory points and concentrations (x_i, y_i, c_i) as calculated above. The Gaussian kernel is used for wind direction since it is defined over an unbounded range, therefore smoothing with the Epannechnikov kernel K(x) is chosen for this work. By definition,

$$K(x) = 0.75\ (1 - x^2) \quad , -1 \le x \le 1, \text{ and 0 otherwise.}$$

Define some equally spaced set of x and y coordinates given by (X_i, Y_j). Then the expectative average concentration E (C|(X_i, Y_j)) at the receptor of air that has passed over point (X_i, Y_j) is given by

$$E(C \mid X_i, Y_j) = \frac{\sum_{k=1}^{N} K\left(\frac{X_j - x_k}{h}\right) K\left(\frac{Y_j - y_k}{h}\right) C_k}{\sum_{k=1}^{N} K\left(\frac{X_j - x_k}{h}\right) K\left(\frac{Y_j - y_k}{h}\right)} \qquad (f)$$

Where h is the smoothing parameter defined by

$$h = \frac{FWHM}{\sqrt{2}}$$

and FWHM is an adjustable parameter giving the full width at half maximum of the smoothing function. The variance of this estimate is given by

$$\mathrm{Var}(E(C \mid X_i, Y_j)) = \|K\|^4 \frac{\sum_{k=1}^{N} K\left(\frac{X_j - x_k}{h}\right) K\left(\frac{Y_j - y_k}{h}\right)(C_k - E(C \mid X_i, Y_j))^2}{\left[\sum_{k=1}^{N} K\left(\frac{X_j - x_k}{h}\right) K\left(\frac{Y_j - y_k}{h}\right)\right]^2} \qquad \text{(g)}$$

where,

$\|K\|^2 = \int_{-\infty}^{\infty} K^2(x)dx = 0.6$, for the Epannechnikov kernel. (Henry, 2008).

Figure 15 is a typical Nonparametric Trajectories Analysis map illustrates SO_2 at North Long Beach site during July 2005, using 2 hr back trajectories from wind speed and direction of multiple sites with HALO 1500m.

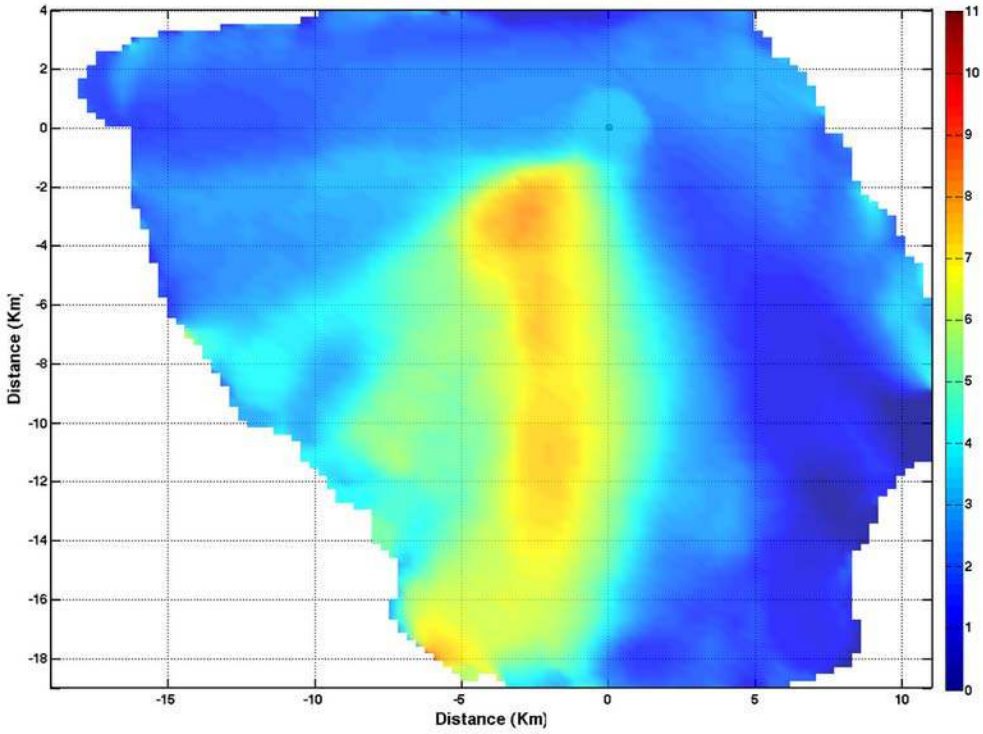


Figure 15 – A NTA graph illustrates SO2 at North Long Beach site during July 2005, using 2 –hr meteorological data of multiple sites to make back trajectories with HALO 1500 m. fwhm = [4 4]

C. Source Apportionment

In order to quantify the expected value of the pollutant, it will start from taking advantage from the basic nonparametric back trajectory analysis shown in equation (f) and making point source response test to estimate the contribution of each sector.

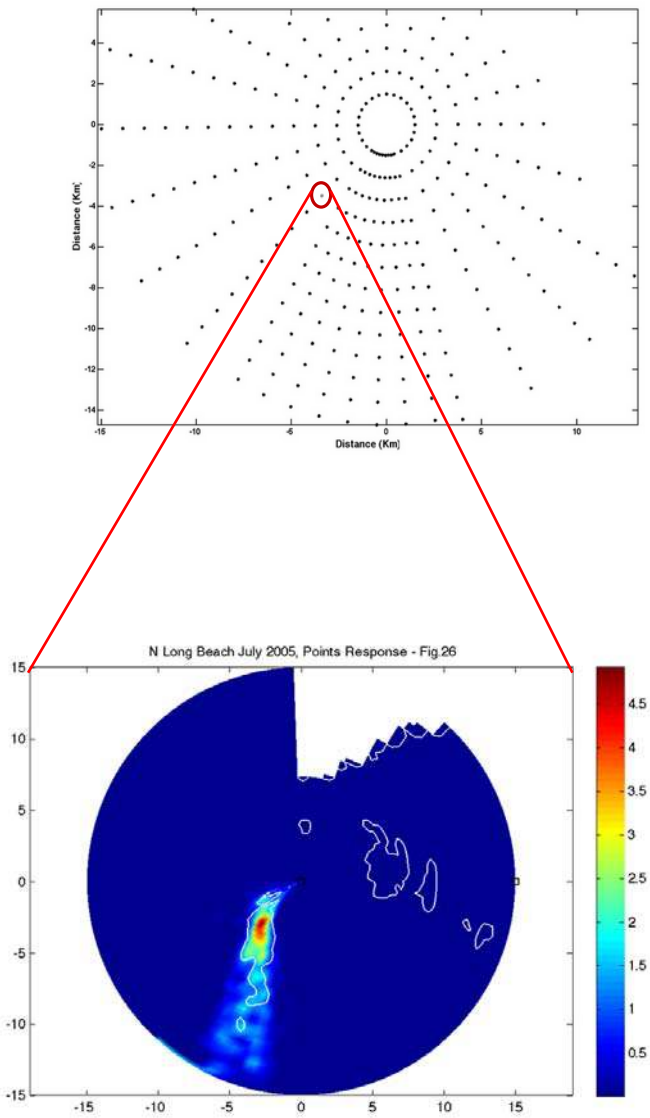
a. Point Source Response Method

The basic concept of Point Source Response (PSR) method is that the expected average concentration map can be linearly combined by a series point source response results. The following flow charts Figure 16 and Figure 17 express the overview of the main idea of point source response technique.

PSR technique is a very useful method to quantify the contribution and effect of vary local sources using the response to a single point source to the whole system. Using this method to analyze data should be careful and consider the following conditions:

1. The point source response can be estimated from the existing concentration data and trajectories.
2. The NTA map of real sources is a linear combination of PSR functions.

Step 1:



Set up several source possible test points around the graph range.

Make point source response map for each of these test point.

Figure 16 – Express the basic idea of making point source response.

Step 2:

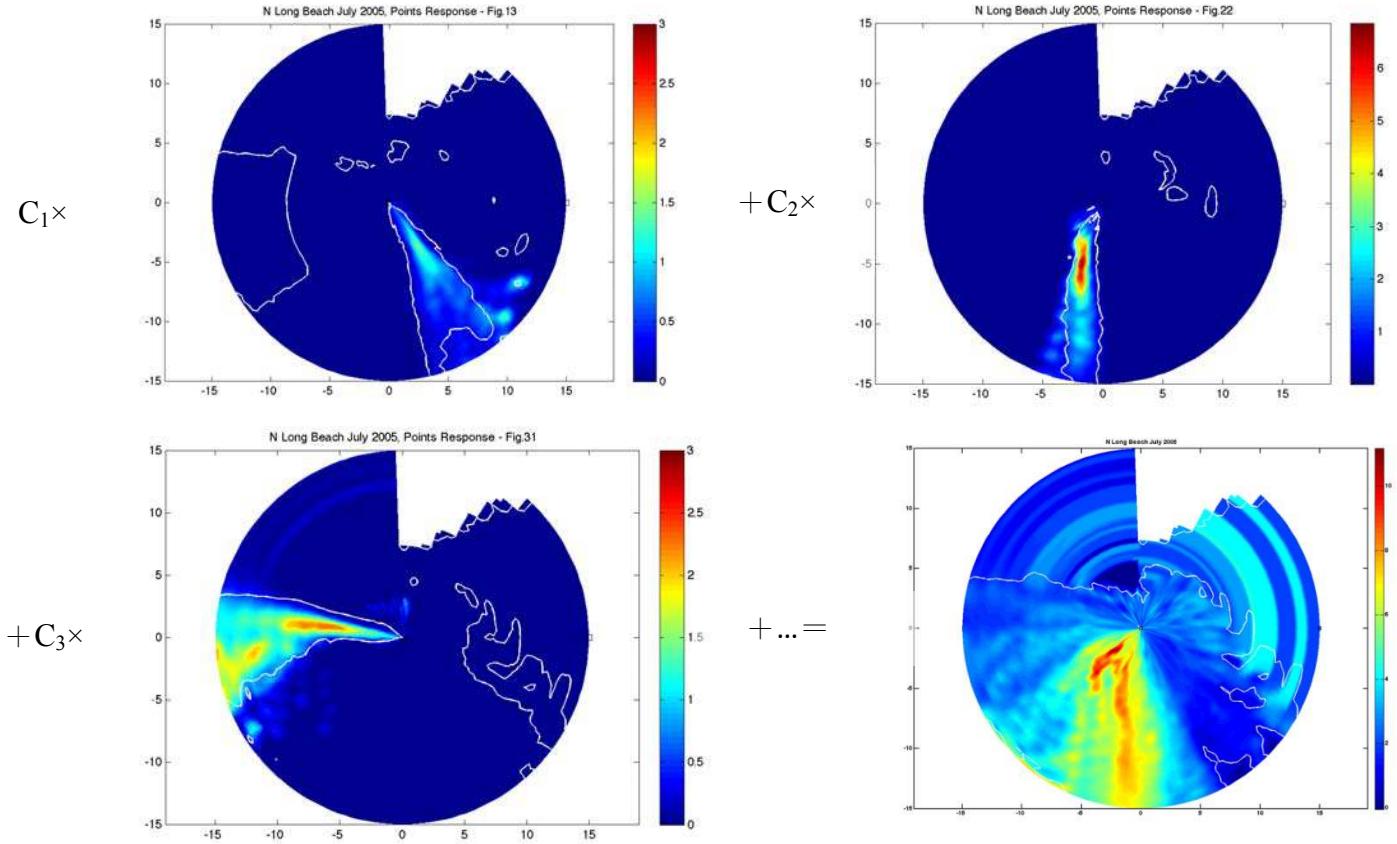


Figure 17 – A NTA graph illustrates SO_2 at North Long Beach site during July 2005, using 2 –hr meteorological data of multiple sites to make back trajectories with HALO 1500 m.

Combine several point source response maps can be made up a simulated average concentration profile. By comparing this simulated map with the expected average concentration map, the weighting coefficients, a series of C_k, can be estimated.

b. Point Source Response Mathematics Methodology

After understanding the basic idea of point source response technique, this section mentions about the detailed mathematics methodology.

The first step to start PSR method is to make a new point source concentration from the original observed concentration data.

$$C_{PR,k} = C_{obs} \times Q_k \qquad \text{(h)}$$

Where $C_{PR,k}$ is the kth new point source response concentration;

C_{obs} is the original observed concentration;

Q_k is the kth Gaussian decay factor, which is defined as

$Q_k = \exp[-\frac{1}{2}(\frac{(X - x_k)^2 + (Y - y_k)^2}{\sigma^2})]$, X and Y are the points of back trajectories, and x_k, y_k defined as the location of the kth test point source, and σ is taken to be the full width at half maximum need in the NTA. Note choosing test point source has some criteria:

1. Two pairs of (x_k, y_k) should not be too close to each other, and the intervals should be at least larger than smoothing parameter h in equation (f). Otherwise, point sources will be too close to distinguished.
2. Point sources selected should avoid of points nearby the original point since all trajectories move toward the original point, if the chosen points were too close to (0,0), after averaging by nonparametric regression these selected points will all show the same result.

By inputting the new point source response concentration, the expected average concentration profiles of a pollutant is known as:

$$E_k(C_{PR} \mid X_i, Y_j) = \frac{\sum_{m=1}^{N} K\left(\frac{X_j - x_m}{h}\right) K\left(\frac{Y_j - y_m}{h}\right) C_{PR,k}^{m}}{\sum_{k=1}^{N} K\left(\frac{X_j - x_m}{h}\right) K\left(\frac{Y_j - y_m}{h}\right)} \qquad \text{(i)}$$

K(x) is Epannechnikov kernel defined as

$$K(x) = 0.75\,(1 - x^2) \qquad -1 \le x \le 1, \text{ and 0 otherwise.}$$

$E_k(C_{PR} \mid X_i, Y_j)$ is the expected value of the kth point source response; x_m, and y_m are the mth trajectory points set, and $C_{PR,k}^{m}$ means the mth concentration of $C_{PR,k}$ their correspondent new point source concentration.

c. Principal Components Regression

The second step is to find the expected value of observed concentration as a linear combination of the results of point source response. The expected average concentration map can be the sum of point source response maps multiplied by a weighting coefficient. The relationship is shown as below.

$$E(C_{obs} \mid X_i, Y_j) = \sum_{k=1}^{N} C_k \times E_k(C_{PR} \mid X_i, Y_j) \qquad \text{(j)}$$

$E(C_{obs} \mid X_i, Y_j)$ is the expected value of observed concentration associated with back trajectory points, and C_k is the weighting coefficient for each point source response test. Physically, $C_k \times E_k(C_{PR} \mid X_i, Y_j)$ represent the contribution of a source k at (x_k, y_k) on the receptor.

In order to find the weighting coefficient C_k, the above mathematic relation equation (j) can be rewritten as a matrix format shown as equation (k).

$$\mathbf{M} = \mathbf{P} \times \mathbf{C} \qquad \text{(k)}$$

Where, **M** is the expected average concentration vector for real data with size (m × 1);

P is a matrix with all point source response maps, which is a (m × n) matrix;

C is the weighting coefficient vector with size (n × 1).

According to the basic matrices theory to determine **C** component, it is expressed as

$$\mathbf{C} = (\mathbf{P}^{\mathbf{T}} \times \mathbf{P})^{-1} \times \mathbf{P}^{\mathbf{T}} \times \mathbf{M} \qquad (l)$$

And $(\mathbf{P}^{\mathbf{T}} \times \mathbf{P}) = \mathbf{P}^{\mathbf{2}}$

However, since some of the point sources are too close to each other, these PSR results should be very linearly similar. Hence, P^2 as a combination of all PRS maps, there will be too many very small values to calculate the inverse matrix directly. Therefore, before we operate this nearly singular matrix, a step called "Singular Value Decomposition (SVD)" can help to solve this difficulty.

The principle of SVD bases on Principal Components Regression (PCR), it seeks to decompose the singular matrix to several related sub matrices to avoid of the singular values in the matrix and help to recreate an inversable matrix.

According to SVD, the square matrix P^2 can be rewritten as equation (m) and its inverse matrix can be written as equation (n).

$$\mathbf{P}^{\mathbf{2}} = \mathbf{U} \times \mathbf{S} \times \mathbf{V}^{\mathbf{T}} \qquad (m)$$

$$(\mathbf{P}^{\mathbf{2}})^{-1} = \mathbf{V} \times \mathbf{S}^{-1} \times \mathbf{U}^{\mathbf{T}} \qquad (n)$$

Where, $\mathbf{P^2}$ is an (n × n) square matrix, **U** is a (n × k) matrix, **S** is a (k ×k) diagonal matrix, and **V** is a (n × k) matrix.

Thus, by choosing a proper **S** to find a similar $\mathbf{P^{2\prime}}$ and consider equation (l) and (n), vector **C** can be determined. The following is a simple example to help understand how to reproduce a proper $\mathbf{P^2}$ matrix.

Example:

$\mathbf{P^2} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}$, obviously, the columns in $\mathbf{P^2}$ are linearly dependent and do not have an inverse matrix.

By considering Singular Value Decomposition, $\mathbf{P^2}$ matrix can be decomposed to the product of **U**, **S** and **V**.

$$\mathbf{U} = \begin{bmatrix} -0.1347 & -0.8257 & 0.3812 & -0.3933 \\ -0.3408 & -0.4288 & -0.2152 & 0.8085 \\ -0.5468 & -0.0319 & -0.7133 & -0.4372 \\ -0.7528 & 0.3650 & 0.5473 & 0.0220 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} 38.6227 & 0 & 0 & 0 \\ 0 & 2.0713 & 0 & 0 \\ 0 & 0 & 10^{-15} & 0 \\ 0 & 0 & 0 & 10^{-16} \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} -0.4284 & 0.7187 & 0.2826 & -0.4692 \\ -0.4744 & 0.2738 & -0.7265 & 0.4150 \\ -0.5203 & -0.1710 & 0.6053 & 0.5776 \\ -0.5663 & -0.6159 & -0.1614 & -0.5234 \end{bmatrix}$$

By skipping the very small value in diagonal matrix S, and it can be rewritten as

$$\mathbf{S'} = \begin{bmatrix} 38.6227 & 0 \\ 0 & 2.0713 \end{bmatrix}$$

Theoretically $\mathbf{P^2}$ can be reproduced to a new similar matrix $\mathbf{P^{2'}}$ by $\mathbf{U'} \times \mathbf{S'} \times (\mathbf{V'})^{\mathbf{T}}$

Where

$$\mathbf{U'} = \begin{bmatrix} -0.1347 & -0.8257 \\ -0.3408 & -0.4288 \\ -0.5468 & -0.0319 \\ -0.7528 & 0.3650 \end{bmatrix}$$

$$\mathbf{S'} = \begin{bmatrix} 38.6227 & 0 \\ 0 & 2.0713 \end{bmatrix}$$

$$\mathbf{V'} = \begin{bmatrix} -0.4284 & 0.7187 \\ -0.4744 & 0.2738 \\ -0.5203 & -0.1710 \\ -0.5663 & -0.6159 \end{bmatrix}$$

Therefore, the $\mathbf{P}^{2'} = \begin{bmatrix} 1.000 & 2.000 & 3.000 & 4.000 \\ 4.999 & 5.999 & 6.999 & 8.000 \\ 8.999 & 9.999 & 10.999 & 12.000 \\ 12.999 & 14.000 & 15.000 & 16.000 \end{bmatrix}$

Since $\mathbf{P^{2'}}$ is very similar to $\mathbf{P^2}$, technically these two matrices can be seen as equal.

Hence, the inverse matrix $(\mathbf{P}^2)^{-1} = \begin{bmatrix} -0.2850 & -0.1450 & -0.0050 & 0.1350 \\ -0.1075 & -0.0525 & 0.0025 & 0.0575 \\ 0.0700 & 0.0400 & 0.0100 & -0.0200 \\ 0.2475 & 0.1325 & 0.0175 & -0.0975 \end{bmatrix}$ can be

produced by equation (n), and vector **C** can be found by equation (l).

The SVD method can be applied to a very complicated and large matrix, theoretically, the increase the size of diagonal matrix **S'**, the closer to the real $\mathbf{P^2}$ by product of the three sub matrices can be made. However, as **S'** size increasing, the probability of the components in C vector to become less than zero is also increasing. It should be careful that since **C** vector is a set of weighting coefficients, it should not be less than zero. This is an important criteria when choose the size of **S'**.

Finally, reproducing a NTA map by linearly combining PSR and comparing the NTA map of real data, a best fit of PSR combination can be determined.

d. Source Apportionment

The final step is to estimate the fraction of each point source response result.

$$S_k(x_k, y_k) = \frac{C_k \times E_k(C_{PR} \mid X_i, Y_j)}{\sum_{k=1}^{N} C_k \times E_k(C_{PR} \mid X_i, Y_j)} \qquad (o)$$

S_k is the fraction of each point source. Since C_k is the weighting coefficient for each point source response, $C_k \times E_k (C_{PR} \mid X_i, Y_j)$ reflects the average contribution for each point source. Therefore, S_k can be expressed as the apportionment of each source. When the coordinates system changes from Cartesians coordinates to polar coordinates, equation (k) can also be rewritten as the following (r, θ) form.

$$S_k(r_k, \theta_k) = \frac{C_k \times E_k(C_{PR} \mid r_i, \theta_j)}{\sum_{k=1}^{N} C_k \times E_k(C_{PR} \mid r_i, \theta_j)} \qquad \text{(p)}$$

D. Data Screening

Data used for this research were reviewed regularly for their quality. When data were obtained from the SCAQMD, the numbers were partially quality controlled (either by software integrated in the monitoring instrument or by SCAQMD staffs) if there was a known period of bad data due to instrument malfunction or an instrument going offline for repair or scheduled maintenance. Despite the steps taken at SCAQMD and because

bad data were passed through the initial quality control steps, the data file had to be closely examined again. Even after these initial efforts, bad data had to be removed as they were spotted during the analysis. Once these results were produced, their validity was back-traced with the data.

The main mechanism of quality control for the data from the initial package during this study was a line-by-line inspection of numbers in the data sheet. Data files were initially in simple text file format. Due to their large size, these files needed to be read by more advanced text editor software such as Note Tab Light or worked by Unix GREP and Regular Expression patterns. The numbers were grouped monthly, seasonally, or for a specific period of time for analysis. Then these numbers were manually inspected line by line or plotted in a “simple concentration vs. time” chart to visually identify any obvious strange pattern or outliers. One example of strange pattern was very high concentrations during 3 – 4 AM on daily basis which was due to instruments’ automatic self-calibration. Figure 18 shows the pattern of this routinely high concentration occurrence as well as other occasionally high concentration

occurrence. Figure 19 shows the pattern of SO_2 concentration after removing the routinely high values and other occasional outliers. The software of the monitoring instrument already tagged some invalid data automatically. The value of -99 were inputted whenever the instrument automatically determined that the measurement value was missing or over ranged. Since these numbers were consistent throughout all the data files, they were easily replaced as NaN (Not a Number) before being fed to data analysis by the Matlab application.

For PM_{10}, any numbers above 500 $\mu g/m^3$ were considered unreasonable outliers and would be eliminated. From the experience of colleagues and professors', it was determined that any PM_{10}values over 200 $\mu g/m^3$ in the Los Angeles air basin was questionable. However, in order to avoid subjective bias and because of the relatively scarce existence of such values and the robustness of nonparametric regression analysis, most of the PM_{10}values ranging 150 – 250 $\mu g/m^3$ were not removed. Accordingly, for SO_2concentrations, values above 100 ppb were eliminated from the data series (Yoon, 2004).

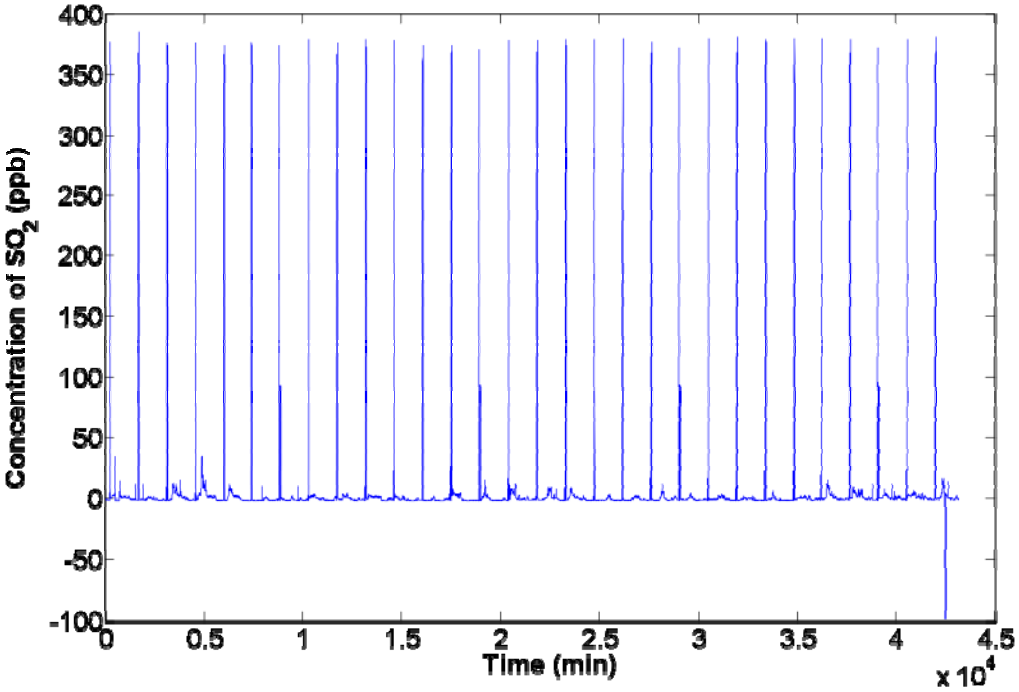


Figure 18 - SO_2 concentration before removing extraordinary high concentrations for North Long Beach site during June 2006.

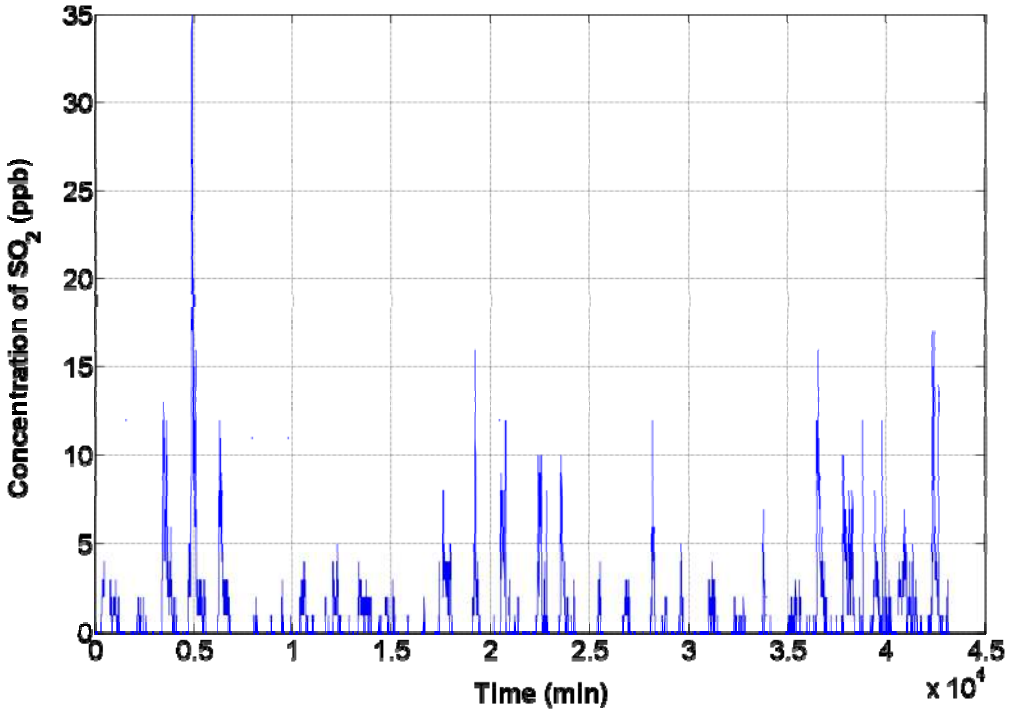


Figure 19 - SO_2 concentration after removing extraordinary high values and other occasional outliers for North Long Beach site during June 2006.

E. Limitations of the Study

This study focuses on data analysis of data collected by air pollution agencies. Therefore, there will be no emphasis on the methods of collecting the data and details of measuring the concentrations and monitoring instruments. In addition, identifying

physical or chemical composition of pollutants is not part of this research. This proposed research will remain focused on the analytical methods for identifying and quantifying the sources of emissions by using the data gathered and distributed by air regulatory agencies. In order for that, it has been assumed the monitors are good indicator and representative of the pollutants within the area around them regardless of the distance and considering the terrain. This assumption has some downsides to it. The monitors have more sensitivity to closer emission sources. The further the sources are located from the monitor; it is less likely that concentrations be measured by the monitors. On the other hand, in order for pollutions to reach the monitor, the sources should be in the direction of prevailing wind. If not on the direction of prevailing winds, the pollutants either will not be measured by monitors or not enough data will be available to run the trajectory analysis.

The other downside to this method is the fact that the further the sources from the monitor, the more distant points will be for the calculations, so it decreases the accuracy of the method. It will be more accurate to calculate the sources with close points to each other comparing to more scattered ones. On the other hand, the

measurement of wind speed and wind direction may not be accurate because of the presence of obstruction to the airflow. As mentioned before, there presence of a 10 story building about 120 m to the northwest of the North Long Beach monitoring site caused the trajectories to the north and west to be unreliable. That is why only the results given do not rely on trajectories from northwest of the monitor (Henry, 2006). In order to have reliable results, it is necessary to have sufficient data. The data used for this research are during January, February, and March of 2006, and April, July, and October of 2007 measurement for every one minute which is about 130,000 data points. In order to speed the computing the trajectories were used to for every 10 minutes instead of every minute. Even by doing so, there is still enough data to analyze with no significantly different results. Thus, if 1-mintue data are available, the method could be easily applied to monthly average or even as little as data for one or two days. The main limitation here is not the number of data points, but the fact that for a short time period such as one or two days it is unlikely that there will be sufficient variation in the wind to produce trajectories that pass over the sources of interest. Thus, the minimum number of days that the method can be applied to is determined by the variability of

the wind and the likelihood that the sources of interest will impact the monitor during the period (Henry, 2006).

Chapter 5

Results and Discussion

The basic assumption of Point Source Response (PRS) method to estimate source apportionments of pollution from different sources is that the map of Nonparametric Trajectory Analysis (NTA) of real data can be linearly combined by a series point source response results.

In order to have a better understanding the source apportionments results by PSR, this chapter follows the steps of data analyzed methodology from the previous chapter: first, discuss backtrajectory and NTA, and second apply PSR to define the response of test point sources. Finally, analyze PSR results by Principal Components Regression (PCR)

to make the best fits NTA map of concentration, and estimate the possible apportionments of sources. In addition, to describe trajectories more accurate, polar coordinate plots will be used to express the NTA map in this discussion instead of using rectangular system.

The parameters and objectives of this study were using 1- minute wind speed and direction data and using observed concentrations of PM_{10} and SO_2 for North Long Beach site and Rubidoux site respectively. To achieve better results, calculating back trajectories in multi sites was only considered.

A. North Long Beach SO_2

a. Trajectory and NTA

In order to have an accurate trajectory, it is not necessary to consider too many sites that too far away from North Long Beach. There are only five nearby monitoring sites

were considered to make back trajectories since the trajectories are only 2 hours and the NTA considered range is in 15 km radius. The sites for North Long Beach are Los Angeles International Air Port (LAX), West Los Angeles, Lynwood, South Long Beach, and North Long Beach. Figure 20 shows the trajectories that we used in this study

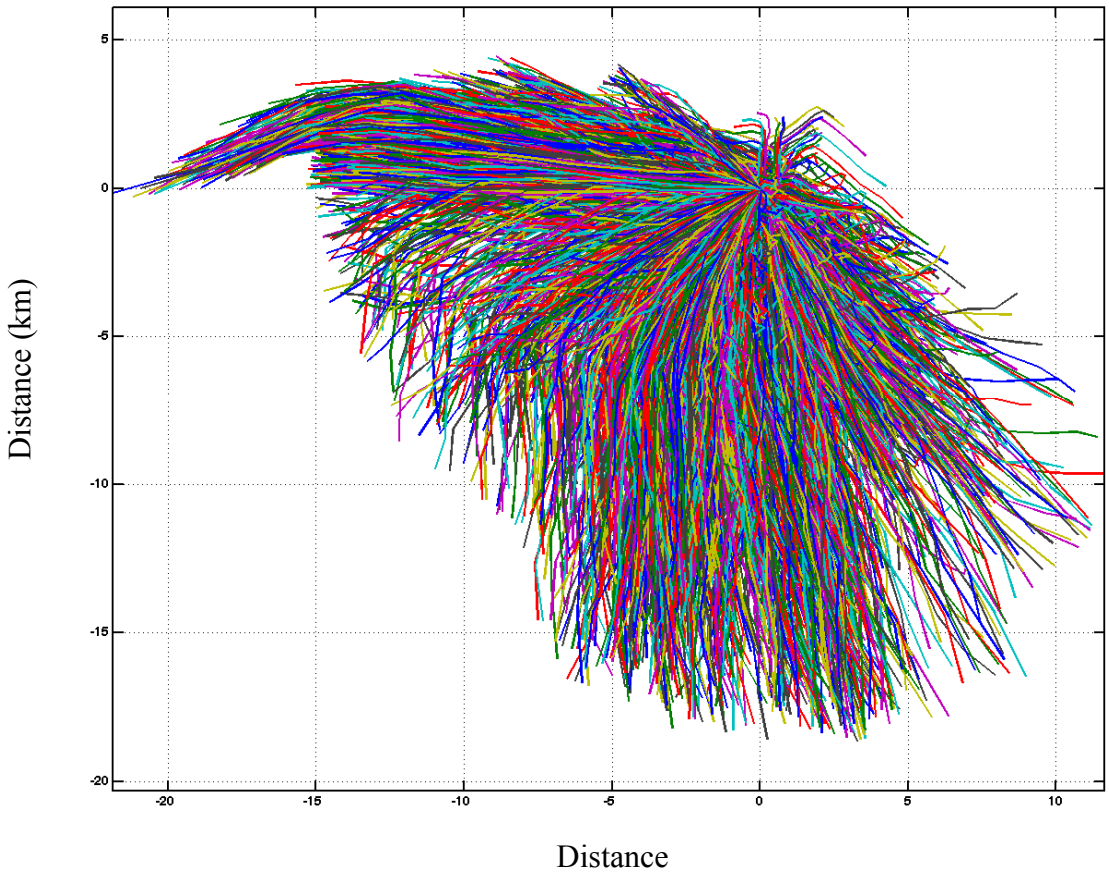


Figure 20 – 2 hours trajectories made by meteorological data from five nearby monitoring stations for North Long Beach in July 2005. The HALO range is 1500 m.

After associating SO_2 concentration with trajectories results by NTA, an NTA map of average concentration can be made. Figure 21 shows the NTA map for SO_2 approximated distribution at North Long Beach.

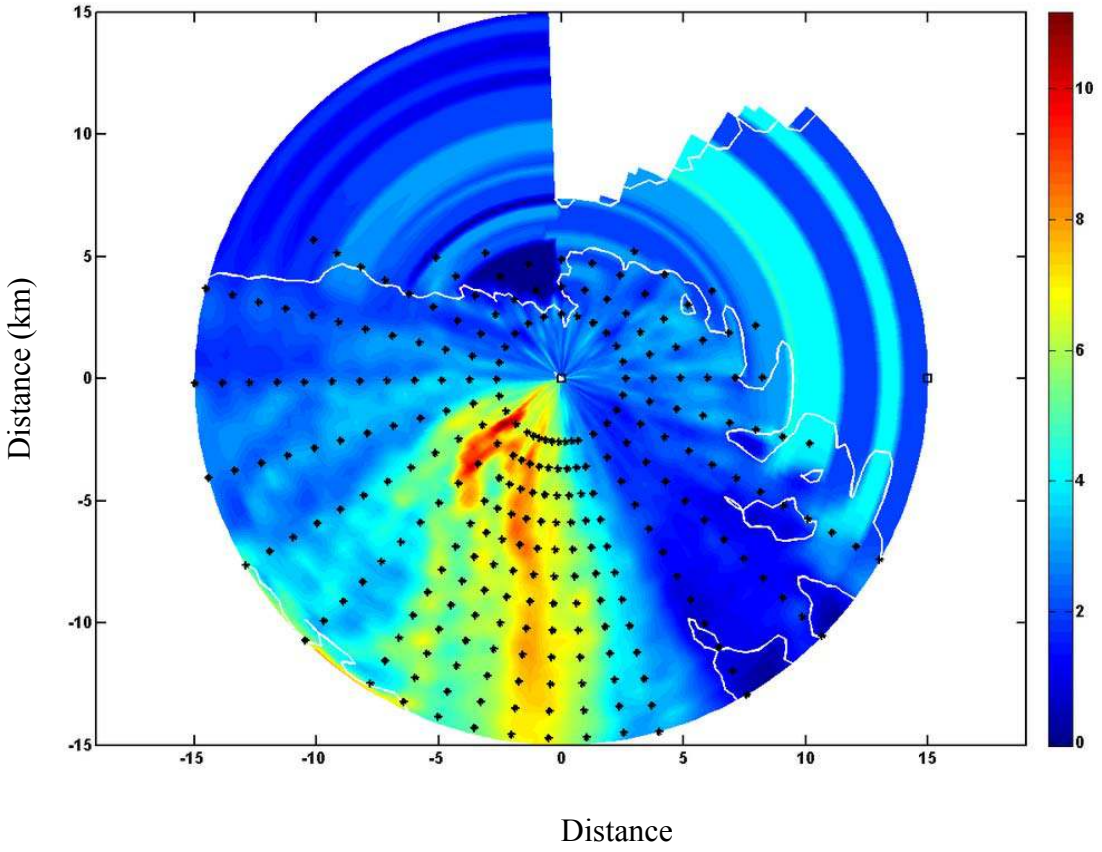


Figure 21 – NTA map of real concentration data of SO_2 at North Long Beach during July 2005. The black points were test point sources fwhm = [5 1]

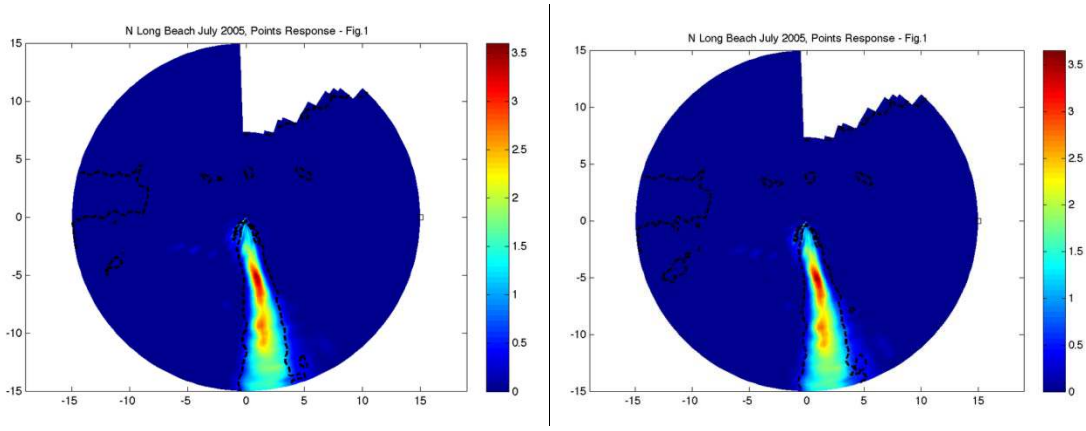
b. Point Source Response Analysis

Before starting Point Source Response analysis, point source should be designed. The black points shown in figure 21 are the assigned test point sources. Choosing point sources can use the following suggestions:

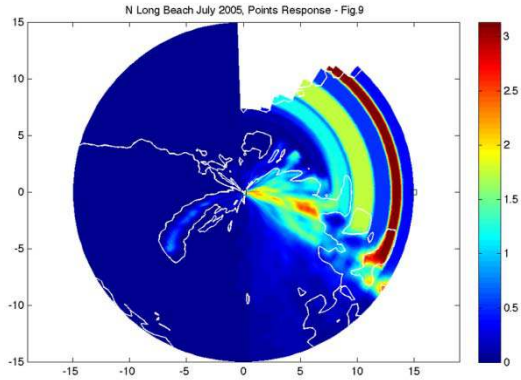
1. Point sources should not be too close to each other, in case of too many overlap of the result since NTA was taking average in its full width at half maximum.
2. It might have big influences when choosing points close to the origin point (receptor), because all trajectories move toward the original points.
3. In order to reduce the noise, test points should avoid of the areas of insufficient data.

Figure 22 shows some of the special cases of choosing point sources. Case (a) shows two point sources are too close to distinguish the different. Case (b) shows point source is too close to the insufficient data area, so that the point source response with some noise. Case (c) shows point source is too close the original point; therefore, the real source can barely show out.

(a) Two nearby point sources



(b) Close to area of insufficient data



(c) Close to the origin point

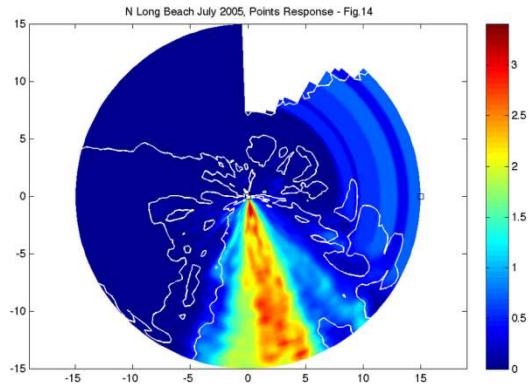


Figure 22 – Different conditions of point source responses

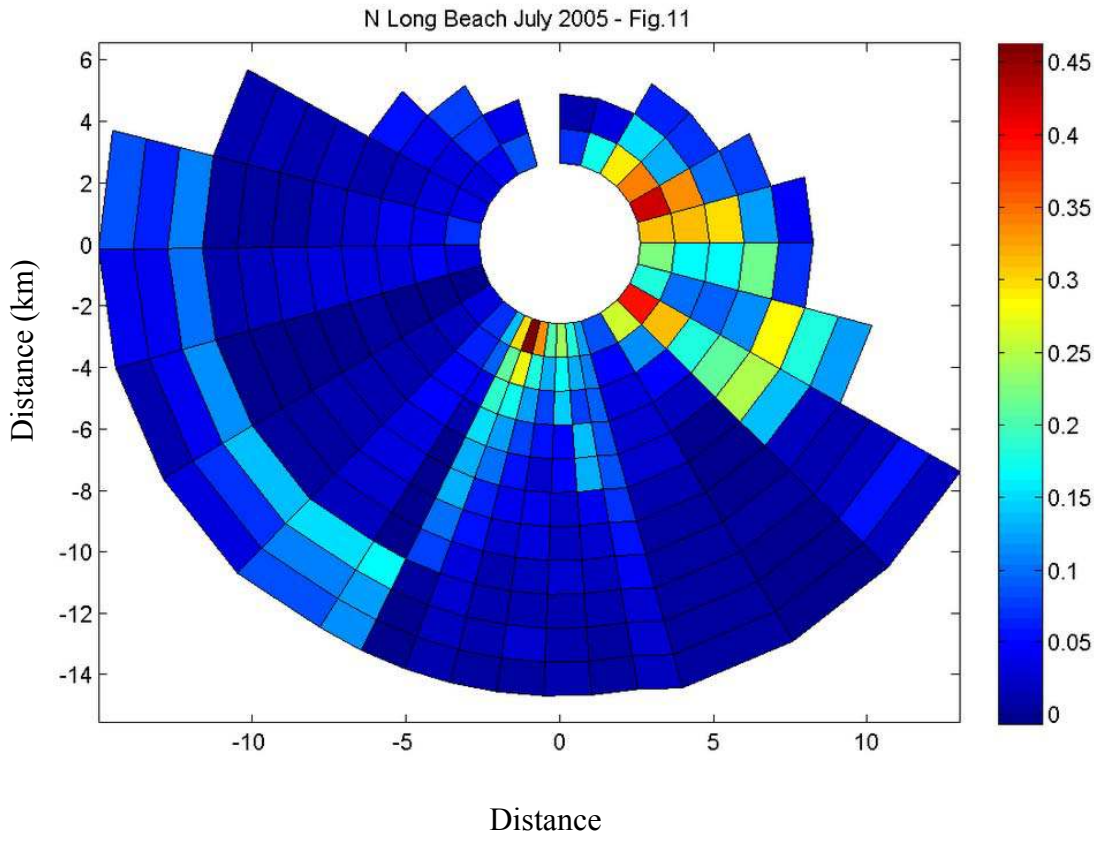


Figure 23 – Choosing weighting coefficients C_k at the first 11 Eigenvalues of point source responses

According to Principal Components Regression (PCR), the weighting coefficient, C_k can be determined. However, in order to find the best fits of PSR result, we should check each set of C_k, and reproduce the NTA map for average concentration from the linearly weight sum of PRS. The result shows that choosing the first 11 Eigenvalues of the PRS to reproduce NTA map can have the best fits NTA result.

Figure 23 shows the distribution of the 11th set of weighting coefficients C_k for each point source response and Figure 24 shows the result of reproducing NTA map by linearly weight sum of PSR.

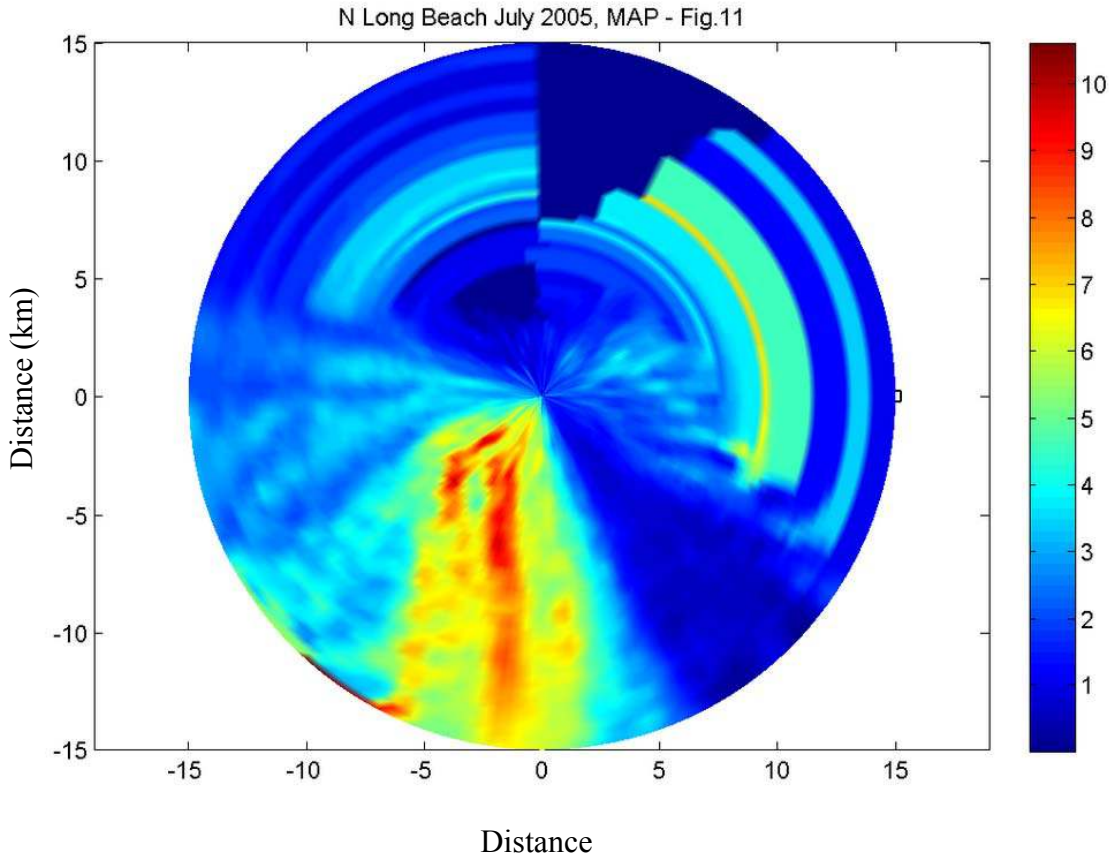


Figure 24 – Reproducing the best fits of NTA map by weight sum of PSR

c. Source apportionments estimate

Since the best fits of NTA map of concentration can be reproduced by linearly weight sum of PSR, PSR can be further analyzed to estimate the source apportionments for sources of pollution. Using the equation (o) in previous chapter,

$$S_k(x_k, y_k) = \frac{C_k \times E_k(C_{PR} \mid X_i, Y_j)}{\sum_{k=1}^{N} C_k \times E_k(C_{PR} \mid X_i, Y_j)}$$

the total distribution of source apportionments can be determined and plot in Figure 25. The graph shows that the pollutant sources mostly were from the south, and distributed around 170 to 210 degrees of Azimuth. In addition, most of the sources are close to the receptor, and for these sources are away from the receptor, longer than 10 km, they may have too many uncertain noises to consider as sources. The detailed results are plot in Figure 26 and Table 2 to illustrate some specific highly pollution sources angles and the percentage of concentration apportionments along with the distance from the receptor. For example, at Azimuth angle 115, 130, 182, 195, 210, and 245.

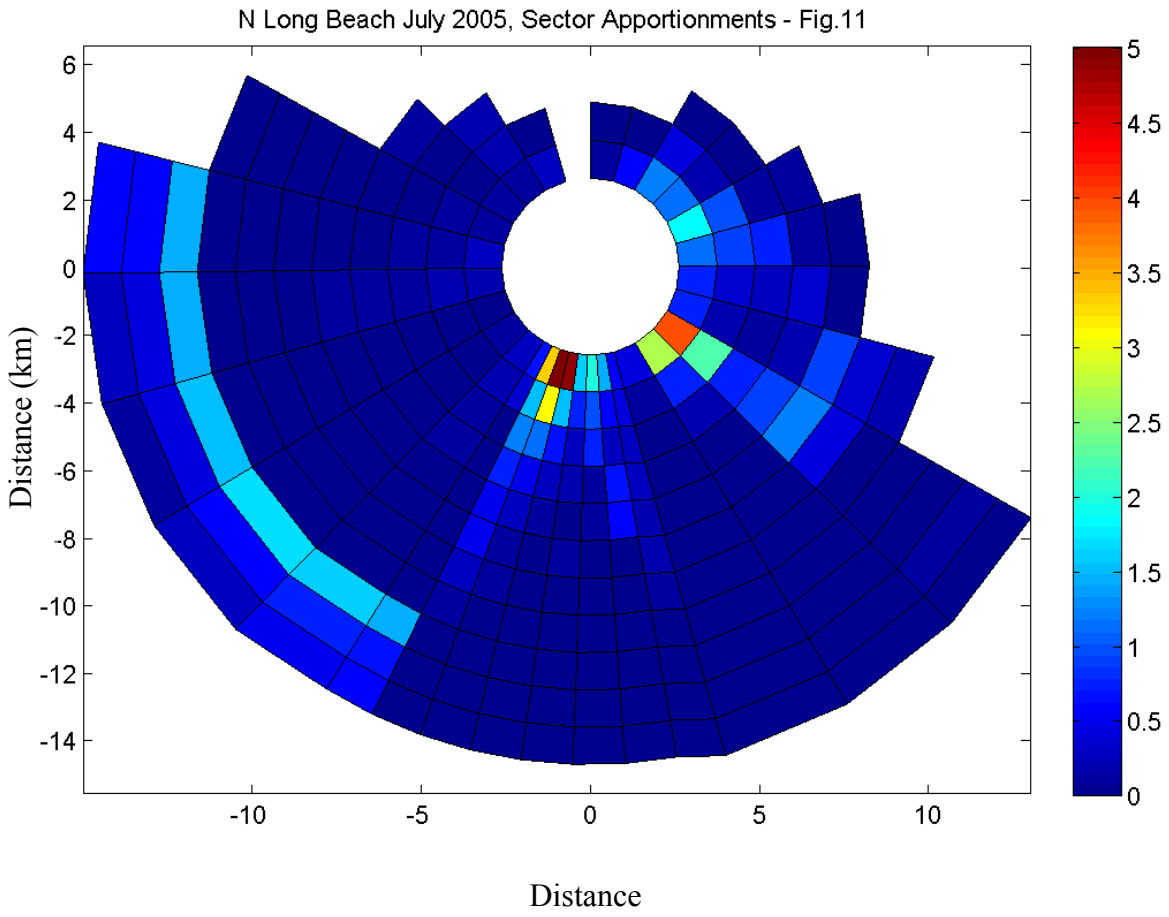
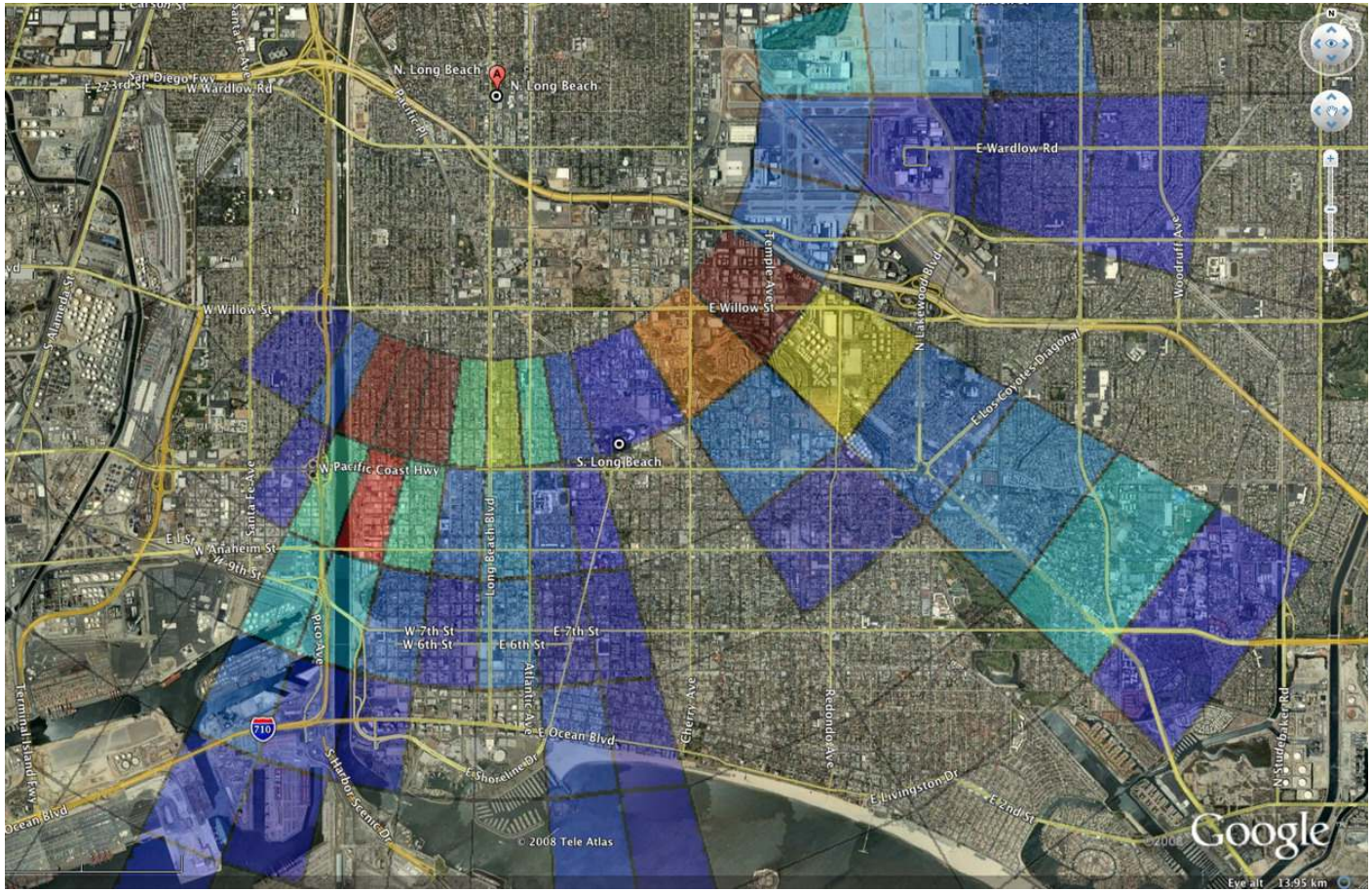


Figure 25 – Choosing weighting coefficients C_k at the first 11 Eigenvalues of point source responses to estimate the sector apportionments for North Long Beach site

Table 2 — Sector apportionment for North Long Beach July 2005

Sector Center (Azimuth)	**Sector Apportionment percentage along with the Radius from the receptor N. Long Beach site (km)**												
	2.6	**3.75**	**4.87**	**6**	**7.12**	**8.25**	**9.37**	**10.5**	**11.62**	**12.75**	**13.87**	**15**	∑
115	0.7276	0.1827	0.1352	0.2004	0.8669	0.321	0.2155	0	0	0	0	0	2.64
130	3.9365	2.2256	0.7175	0.8822	1.1959	0.4242	0.0169	0.0232	0.0445	0.1241	0.0142	0	9.60
182	1.6389	0.7404	0.2968	0.1869	0.0427	0.081	0.044	0.0202	0.0082	0.0224	0.002	0.0011	3.08
195	9.004	3.0559	1.1178	0.5057	0.3068	0.1393	0.0925	0.0387	0.023	0.0134	0.0075	0.0047	14.3
210	0.6737	0.3208	0.0768	0.0044	0	0	0	0	1.4763	0.6668	0.5475	0.4244	4.19
245	0.1359	0.0576	0.0362	0.0266	0.0193	0.0155	0.011	0.0062	1.6512	0.6126	0.2694	0.062	2.90
∑	16.2	6.58	2.38	1.81	2.43	0.98	0.38	0.09	3.20	1.44	0.84	0.49	36.7

Distance (km)



Distance

Figure 26 – Identifying the sources of pollution at North Long Beach site

According to the data from Table 2, it can be simply marked some possible sources on the NTA map. Figure 26 identifies that Long Beach airport nearby area at the east side and port area at the south side of Long Beach area were the most possible SO_2 sources, and this result is correspondent with Dr. Pazokifard's study in 2007.

B. Rubidoux for PM_{10}

a. Trajectory and NTA

Using the same analysis method as the previous section, calculating the 2 – hr back trajectories and NTA for PM_{10} concentration data in $\mu m/m^3$ at Rubidoux site were taken during July 2005 and so were the wind speed and directions data. Figure 27 to Figure 28 are the typical back trajectories, averaged concentration map calculated by NTA. The calculations of back trajectories considered the meteorological data of Crestline, Pomona, San Bernardino, Upland and Rubidoux these five monitoring sites.

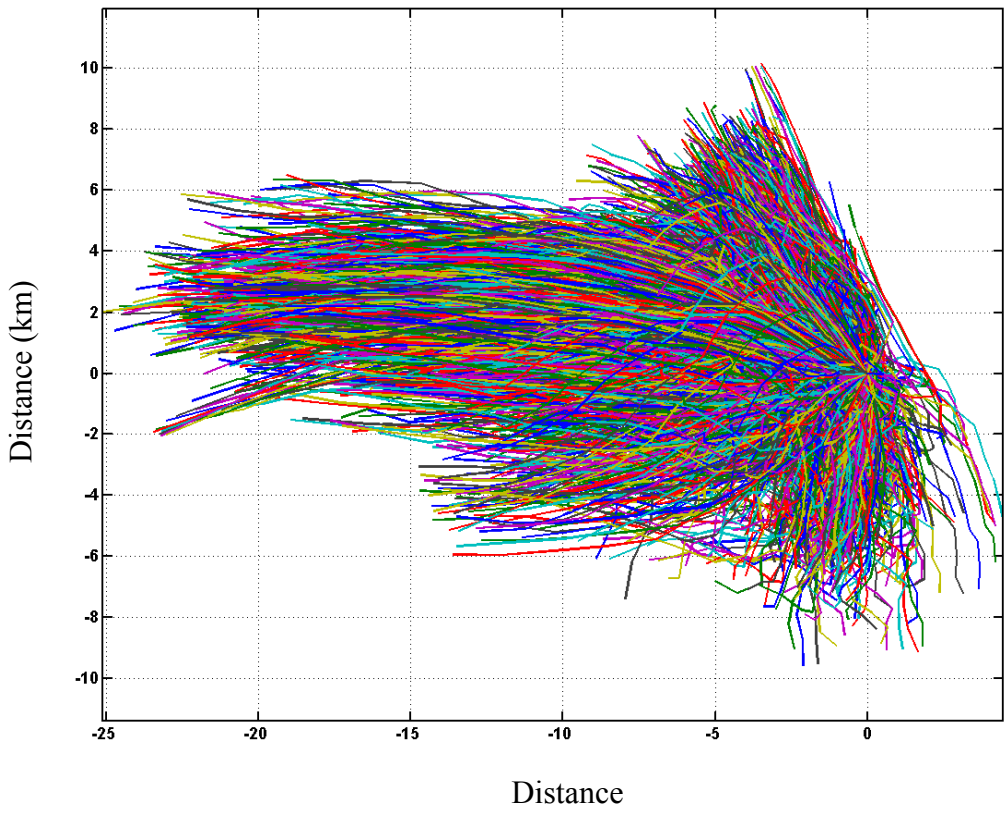


Figure 27 –2 - hour trajectories made by meteorological data from five nearby monitoring stations for Rubidoux in July 2005. The HALO range is 1500 m.

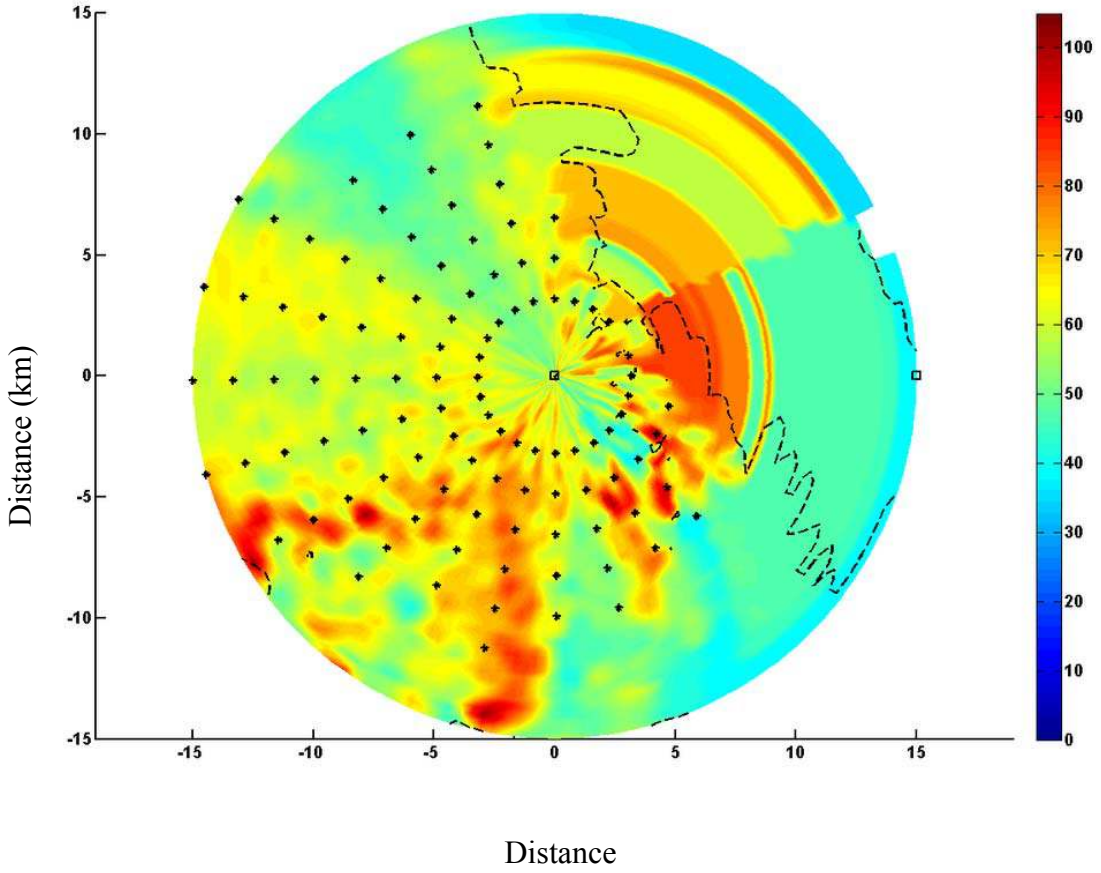


Figure 28 –NTA map of real concentration data of PM_{10} at Rubidoux during July 2005. The black points were test point sources

b. Point Source Response Analysis

Using the same analysis method as the previous section, PCR helps to determine the weighting coefficient, C_k. Therefore, the result shows that choosing the first 12 Eigenvalues of the PRS to reproduce NTA map can have the best fits NTA result.

Figure 29 shows the distribution of the 12th set of weighting coefficients C_k for each point source response and Figure 30 shows the result of reproducing NTA map by linearly weight sum of PSR.

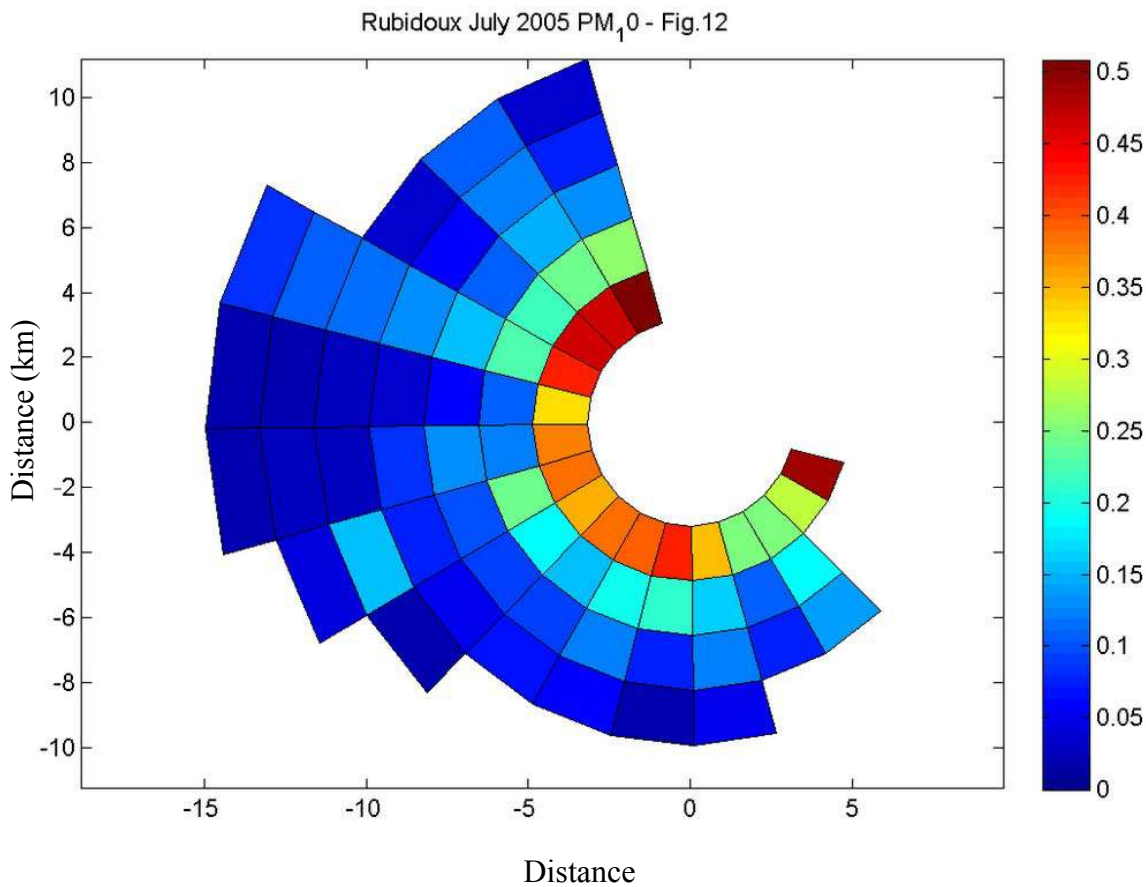


Figure 29 –Choosing weighting coefficients C_k at the first 12 Eigenvalues of point source responses

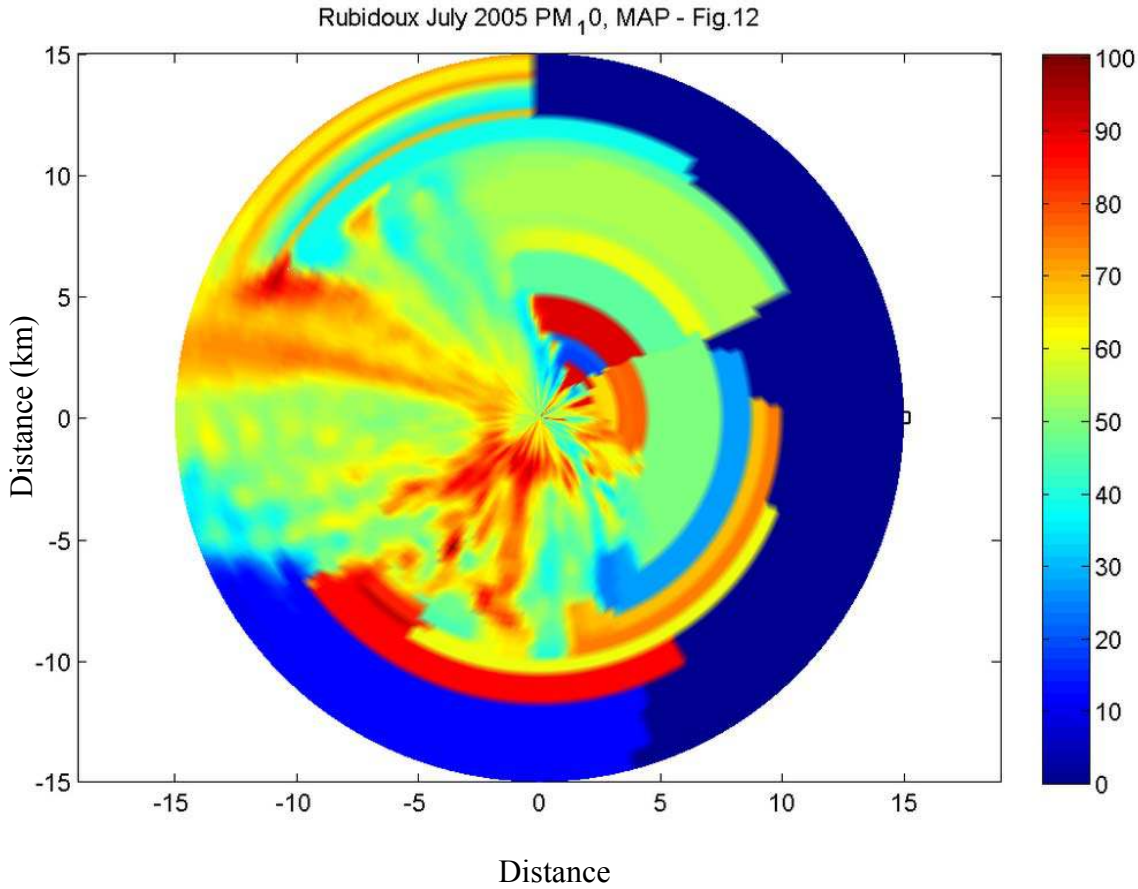


Figure 30 –Reproducing the best fits of NTA map by weight sum of PSR

c. Source apportionments estimate

Following the known step, the total distribution of source apportionments can be determined and plot in Figure 31. Because most of the wind blows from the west side or southwest, there is not enough data to give information regarding the east side.

Therefore, the graph shows most of the pollutant sources from the western, and southwestern. They distributed from 165 and 180 to 270 degrees of Azimuth. In addition, most of the sources were happened around the nearby receptor. Some of the highly possible pollution sources angles were plot in Figure 31 and list in Table 3.

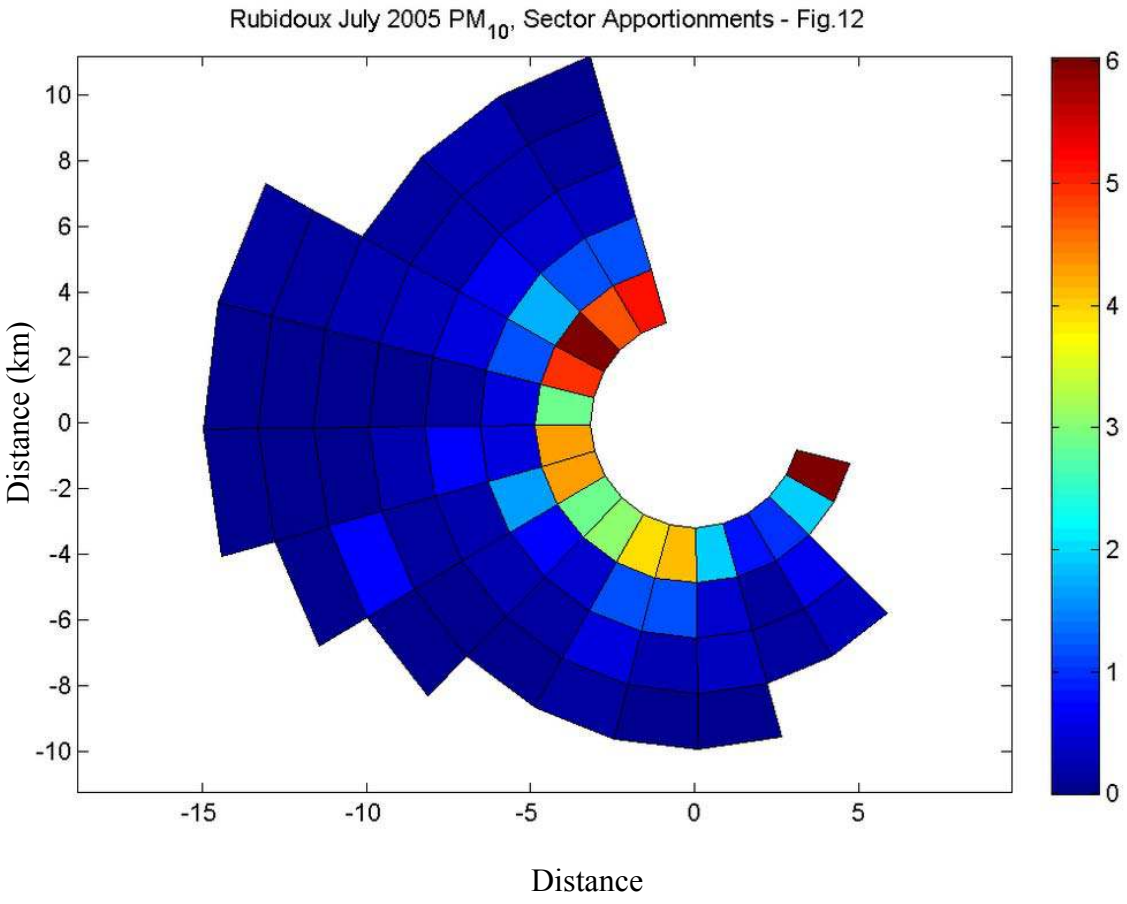
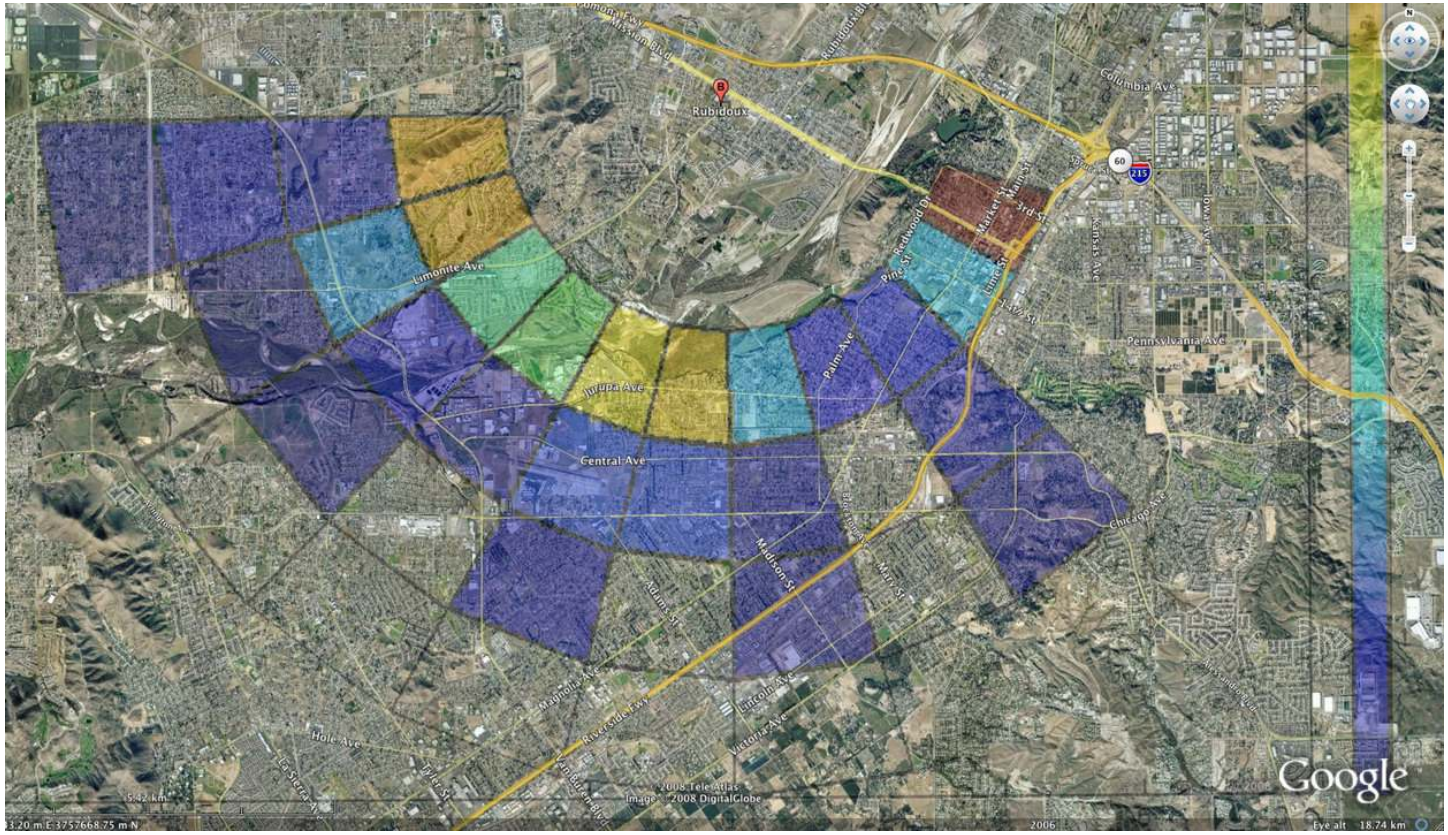


Figure 31 –Choosing weighting coefficients C_k at the first 12 Eigenvalues of point source responses to estimate the sector apportionments for Rubidoux site

Table 3 — Sector apportionment for Rubidoux July 2005

Sector Center (Azimuth)	**Source Apportionment percentage along with the Radius from the receptor Rubidoux site (km)**								
	1.56	**3.25**	**4.95**	**6.62**	**8.31**	**10**	**11.68**	**13.37**	∑
165	0.8290	0.1588	0.0961	0.2205	0.0034	0	0	0	1.3078
195	1.9431	0.4556	0.3188	0.0590	0.0307	0	0	0	2.8072
255	4.3089	1.6893	0.2353	0.1441	0.7212	0.0730	0.0087	0	7.1805
270	4.2966	0.5258	0.6689	0.2429	0.0424	0.0332	0.0158	0.0085	5.8341
∑	11.3776	2.8295	1.3191	0.6665	0.7977	0.1062	0.0245	0.0085	17.1296

Distance (km)



Distance

Figure 32 –Identifying the sources of pollution at Rubidoux site

According to the data from Table 3, some specific pollution sources can be located on the NTA map. From Figure 32 represented the south side of Rubidoux is free way 60 crossing through the residential area, and for the west side is a valley residential area surrounded by three main freeways, Freeway 215, 60, 15 and hills. By conjecture, there might be some fugitive sources, or temporary constructions.

Chapter 6

Conclusion

Point Source Response (PSR) extended by Nonparametric Trajectory Analysis (NTA) was trying to solve the difficulties of NTA since NTA did not always define real local sources. PSR improves not only the identification of sources, but also give us information of source apportionment. However, for identification of sources very close to the receptors, PSR has a limitation to recognize the sources in around 3 km radius. This means NTA takes an average in the range of its smoothing parameter, and if the test point source is too close to the receptor, NTA will take an average including the receptor. Taking average in this rage can cause a significant error. On the other hand, some artificial sources will be created by NTA because of the density of trajectories is getting higher toward the receptor. Therefore, the data cannot show us the accurate

pollution sources in this error range, and the suggestion moderate range for PSR is a round from 5 km to 15 km.

In addition, point source response method has limitation on meteorological data. If the pollutants were regional sources instead of local sources, PSR cannot accurate locate the sources. For example, the result of Rubidoux, since most of the pollutants may not be emitted from local sources; we cannot perfectly reproduce the NTA map and estimate the sources

Bibliography

1. Cohen, M., Artz, R., Draxler, R., Miller, P., Poissant, L., Niemi, D., Ratté, D.,Deslauriers, M., Duval, R., Laurin, R., Slotnick, J., Nettesheim, T.,McDonald, J., (2004). "Modeling the atmospheric transport and depositionof mercury to the Great Lakes." *Environmental Resources* V 3, 247-65.

2. Draxler, R. R., Hess, G. D., (1998). "An overview of the HYSPLIT-4 modeling system for trajectories, dispersion and deposition." Australian Meteorological Magazine, v 47, n 4, 295-308.

3. Escudero, M., Stein, A.; Draxler, R.R., Querol, X., Alastuey, A., Castillo, S., Avila, A., (2006). "Determination of the contribution of northern Africa dust source areas to PM10 concentrations over the central Iberian Peninsula using the Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT) model. "Journal of Geophysical Research-Part D-Atmospheres, v 111, n D6, 15.

4. Fisher, A. L., Parsons, M. C., Roberts, S. E., Shea, P. J., Khan F. L., Husain, T., (2003). "Long-term SO2 dispersion modeling over a coastal region." Environmental Technology v 22, 4831-45.

5. Gebhart, K. A., Schichtel, B. A., Barna, M. G., Malm, W. C., (2006). "Quantitative Back-Trajectory Apportionment of Sources of Particulate Sulfate at Big Bend National Park, TX." Atmospheric Environment v 40, 2823-2834.

6. Gryning, S. E., Schiermeier, F. A., (2001). "Air Pollution Modeling and Its Application XIV." Kluwer Academic Publishers.

7. Härdle, W., (1990). "Applied Nonparametric Regression." Cambridge University Press, Cambridge.

8. Henry, R. C., Turner, J. R., (2007). "Nonparametric Wind Regression." Draft for submission to Environmental Science & Technology.

9. Henry, R. C., (2008). "Locating and Quantifying the Impact of Local Sources of Air Pollution." Atmospheric Environment. v42, 358 - 363.

10. Henry, R. C., (2005 – 2006). Internal Communication and Notes.

11. Henry, R. C., Chang, Y.-S., Spiegelman, C. H., (2002). "Location Nearby Sources of Air Pollution by Nonparametric Regression of Atmospheric Concentration on Wind Direction." Atmospheric Environment v 36, 2237-2244.

12. Henry, R. C., Lewis, C. W., Hopke, P. K., Williamson, H. J., (1984). "Review of Receptor Model Fundamentals." Atmospheric Environment, v 18, n 8, 1507-1515.

13. Isard, S. A., Gage, S. H., Comtois, P., Russo, J. H., (2005). "Principles of the Atmospheric Pathway for Invasive Species Applied to Soybean Rust." BioScience, v55, n10, 851(11).

14. Kim, Eugene; Hopke, Philip K.; Edgerton, Eric S., (2004). "Improving source identification of Atlanta aerosol using temperature resolved carbon fractions in positive matrix factorization." Atmospheric Environment, v 38, n 20, 3349-3362.

15. Lewis, Charles W., Norris, Gary A., Conner, Teri L., Henry, Ronald C.(2003). "Source apportionment of phoenix PM2.5 Aerosol with the Unmix receptor model." Journal of the Air and Waste Management Association, v 53, n 3, 325-338.

16. Lucey, D., Hadjiiski, L., Hopke, P.K., Scudlark, J.R., Church, T. (2001)." Identification of sources of pollutants in precipitation measured at the mid-Atlantic US coast using potential source contribution function (PSCF)." Atmospheric Environment, v 35, n 23, 3979-3986.

17. Pazokifard, Babak (2007). "Ph.D. Dissertation : Locating and Quantifying Sources of Air Pollution by Nonparametric Trajectory Analysis." Department of Civil and Environmental Engineering, University of Southern California

18. Riccio, A., Giunta, G., Chianese, E., (2007). "The application of a trajectory classification procedure to interpret air pollution measurements in the urban area of Naples (Southern Italy)." *Science of Total Environment,* v 376, n 1-3, 198-214.

19. Spiegelman, C., Eun Sug Park, (2007). “A computation saving Jackknifeapproach to receptor model uncertainty statements for serially correlated data.” Chemometrics and Intelligent Laboratory Systems, v 88, n 2, 170-82.

20. Sprent, P., (1993). “Applied Nonparametric Statistical Methods.” Chapman & Hall, London.

21. Srivastava, Anjali, Sengupta, B., Dutta, S.A. (2004). “Source apportionment of ambient VOCs in Delhi. ”Science of the Total Environment, v 343, n 1-3, 207-220.

22. Steven, G. B., Anna, F., Sean, M. R,, Paul, T. R., Hilary, R. H., Darcy, J. A., (2007). “Source apportionment of fine particulate matter in Phoenix, AZ, using positive matrix factorization.” Journal of the Air & Waste Management Association, v 6, 741(12).

23. Sturman, A., Zawar-R., P., (2002). “Application of Back-Trajectory Techniques to the Delimitation of Urban Clean Air Zones.” Atmospheric Environment, v 36, 3339-3350.

24. Subhash, S., Honrath, R. E., (1999). “Back-Trajectory Analysis of Atmospheric Polychlorinated Biphenyl Concentrations over Lake Superior.” Environmental Science and Technology, v 33, 1509-1515.

25. Tsai, Ying I., Chen, Chien-Lung, (2006). “Atmospheric aerosol composition and source apportionments to aerosol in southern Taiwan.” Atmospheric Environment, v 40, n 25, 4751-4763.

26. Yoon, H., (2004). “Ph.D. Dissertation Proposal: Understanding Ultrafine Particle Concentrations in the Urban Ambient Air Using Nonparametric Regression Analysis.” Department of Civil and Environmental Engineering, University of Southern California.

27. Yu, K. N., Cheung, Y. P., Cheung, T., Henry, R. C., (2004). “Identifying the Impact of Large Urban Airport on Local Air Quality by Nonparametric Regression.” Atmospheric Environment, v 38, 4501-4507.

28. Wand, M.P., Jones, M.C., (1995). “Kernel Smoothing.” Chapman and Hill Ltd., New York and London.