Atmospheric Environment 64 (2013) 150–159

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

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

Evaluating meteorological comparability in air quality studies: Classification and regression trees for primary pollutants in California's South Coast Air Basin

W. Choi [a], S.E. Paulson [a], J. Casmassi [b], A.M. Winer [c,*]

[a] Department of Atmospheric and Oceanic Sciences, 405 Hilgard Ave., University of California, Los Angeles, CA 90095-1565, USA
[b] Planning, Rule Development and Area Sources, California South Coast Air Quality Management District, 21865 Copley Drive, Diamond Bar, CA 91765-4178, USA
[c] Environmental Health Sciences Department, School of Public Health, 650 Charles E. Young Drive South, University of California, Los Angeles, CA 90095-1772, USA

HIGHLIGHTS

- Regression trees for primary pollutants were created using a CART model.
- Primary pollutant levels are largely under control of meteorology in the SoCAB.
- The most important meteorological variables are wind speed and geopotential height.
- Spatial variances in pollutants are well correlated with meteorological conditions.
- CART analysis provides an effective tool for meteorological comparability.

ARTICLE INFO

Article history:
Received 4 April 2012
Received in revised form
1 September 2012
Accepted 19 September 2012

Keywords:
Primary pollutants
Meteorological adjustment
Traffic emissions
Meteorological comparisons

ABSTRACT

Meteorology confounds the comparison of air quality data across time and space. This presents challenges, for example, to comparisons of pollutant concentration data obtained with mobile monitoring platforms on different days and/or locations within the same airshed. In part to address this challenge, we employed a classification and regression tree (CART) modeling approach that can serve as a useful and straightforward tool in such air quality studies, to determine the comparability of meteorological conditions between measurement days and locations as well as to compare primary pollutant concentrations corrected by meteorological conditions. Specifically, regression trees were developed to obtain representative concentrations of traffic-related primary air pollutants such as NO_x and CO, based on meteorological conditions for 2007–2009 in the California South Coast Air Basin (SoCAB). The resulting regression trees showed strong correlations between the regression classifications developed for different pollutant metrics, such as daily CO and NO_x maxima, as well as between monitoring sites. For the SoCAB, the most important meteorological parameters controlling primary pollutant concentrations were the mean surface wind speed, geopotential heights at 925 mbar, the upper air north–south pressure gradient, the daily minimum temperature, relative humidity at 1000 mbar, and vertical stability, in approximate order of importance. The value of developing a regression tree for a single season was also explored by performing CART analysis separately on summer data. Although seasonal classifications were similar to those developed from annual data, the standard deviations of the classification groups were somewhat reduced.

1. Introduction

A growing number of epidemiological studies have shown that exposure to fresh vehicular emissions causes adverse human health effects, including asthma, cardiovascular disease, and adverse birth outcomes (Hoek et al., 2010; Penttinen et al., 2001; Pope et al., 2002). Ultrafine particles (UFP), along with other traffic-related pollutants including nitrogen oxides (NO_x), carbon monoxide (CO), and various organic gases emitted near major roads, are of particular interest in metropolitan areas, including the California South Coast Air Basin (SoCAB). Despite enormous progress in reducing air pollution over the past four decades, the SoCAB remains one of the most polluted regions in the U.S., with mobile sources accounting for 93% and 89% of the total annual emissions of CO and NO_x, respectively, as of 2008 (CARB, 2009).

* Corresponding author. Tel.: +1 310 206 4442; fax: +1 310 206 3358.
E-mail addresses: wschoi@atmos.ucla.edu (W. Choi), paulson@atmos.ucla.edu (S.E. Paulson), jcassmassi@aqmd.gov (J. Casmassi), amwiner@ucla.edu (A.M. Winer).

1352-2310/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atmosenv.2012.09.049

Numerous air quality studies have been conducted near major roads and freeways in the SoCAB (e.g., Hu et al., 2009; Kozawa et al., 2009; Moore et al., 2009; Zhang et al., 2004; Zhu et al., 2006). Because traffic-related pollutants are dependent on meteorological conditions as well as emission rates, the atmospheric concentrations of these pollutants vary from day to day and by location, showing significant heterogeneity in temporal and spatial distributions (Krudysz et al., 2009; Moore et al., 2009; Turner and Allen, 2008). Thus, having an ability to correct for time-variant differences in meteorology for pollutant data for different locations is highly desirable.

To map spatial and temporal variations in pollutant concentrations at high resolution over a large area such as the SoCAB is challenging. In part for this reason, interest in making measurements with instrumented mobile monitoring platforms (MMPs) has been growing recently as high time-resolution instrument capabilities have developed (Hu et al., 2009, 2012; Kozawa et al., 2009). However, because of the high cost of an electrical vehicle fully equipped with sophisticated monitoring instruments as well as staffing costs, making simultaneous measurements with multiple MMPs in more than one area of the same airshed is very expensive. To our knowledge such simultaneous measurements with MMPs have not been reported in the literature. What is required is a reasonably quantitative, systematic and straightforward method to classify the degree of similarity or difference of meteorological conditions between days or locations.

Numerous efforts to adjust tropospheric ozone trends in urban areas for meteorology have been made since the 1980s using a wide range of statistical methodologies such as linear or non-linear regression approaches, tree-based or stratified model approaches, time-series filtering methods, and extreme value theory (Thompson et al., 2001). Because ozone is produced in the atmosphere through photochemical processes, the major meteorological factors affecting ozone concentrations are different from those for traffic-related primary pollutants such as UFP and CO (Horie, 1988; Thompson et al., 2001).

In contrast to the case for ozone, we are not aware of any statistical methodology studies that have produced systematic assessment criteria for meteorological adjustment of traffic-related primary pollutants. Since the classification and regression tree (CART) approach was first developed in the 1960s (Morgan and Sonquist, 1963), it has been applied to purposes as diverse as remote sensing data processing (Tooke et al., 2009), ecological data analysis (De'ath and Fabricius, 2000), medical causation analysis (Hess et al., 1999), and prediction of daily maximum ozone and $PM_{2.5}$ levels (Dye et al., 2003; Horie, 1988).

Here, we develop an objective classification scheme of meteorological conditions for the SoCAB using the CART method. Although in principle this method has some predictive potential for atmospheric pollutant concentrations, it is not a substitute for more sophisticated and quantitative dispersion models (Venkatram, 2004; Venkatram and Cimorelli, 2007). Here we demonstrate that the CART model offers a straightforward approach to the quantitative classification of meteorological effects on primary pollutant concentration which should facilitate comparisons of traffic-related pollutant concentrations between measurement days and locations in the SoCAB.

2. Classification and regression tree modeling approach and parameters for primary pollutants

Here, we build regression trees using two types of data: pollution data from several South Coast Air Quality Management District (SCAQMD) monitoring stations, and a large amount of regional upper air and surface meteorological data. CART itself is straightforward to run. The results are expected to be applicable to the area under the control of the regional meteorology used to build the model. A separate regression tree would need to be developed for other geographical areas of interest, where the "area" refers to a region with similar general weather patterns.

2.1. Model description

The CART method explains distribution or variation of a target variable using a number of explanatory variables each with a linear or non-linear relationship with the target variable. The basic concept of the CART approach is to make a hierarchy of binary decisions, each of which splits distribution/variation of a target variable into two mutually exclusive branches (groups) based on the explanatory variable/value showing the largest reduction in variations in a target variable after the split. Each split branch is then divided into two sub-branches by the other variables, until a set of terminal nodes (leaves) is reached. The details concerning how to determine the terminal nodes (to prune excessively large numbers of splits) and the theoretical underpinning of the CART approach are found in Breiman et al. (1984) and the supplementary material (S1). A target variable is either categorical (classification trees) or numerical (regression trees), and a number of explanatory variables are also either categorical or numerical. Thus, the CART approach allows complicated links between a target variable and various explanatory variables to be clear, easier to interpret, and quantitatively compared.

2.2. Regional parameters

The SoCAB occupies a coastal plain surrounded by mountains on three sides (the San Gabriel, San Bernardino, and San Jacinto mountains) and includes Los Angeles (LA), Orange, and the western portions of Riverside and San Bernardino Counties. The predominant meteorological conditions in the SoCAB are characterized by mild winds and shallow boundary layer heights capped by low-altitude temperature inversions due to a semi-permanent "Pacific High" pressure cell. Prevailing winds are dominated by diurnal cycles of weak off-shore breezes at night and stronger on-shore sea breezes during the day. The mountain escarpments surrounding the SoCAB enhance the pollutant-capping effects, preventing air ventilation (SCAQMD, 2012). Less common weather patterns, occurring primarily in the winter, include storm fronts arriving largely from the north and west, and dry winds arising from high deserts to the east. The latter are referred to as Santa Ana winds.

In this study, the downtown LA (DTLA) monitoring site (N. Main St., 25 km from the coast, 34.07°N/118.23°W) was selected as a representative station to create and investigate the regression trees for traffic-related primary pollutants. Five additional monitoring sites were chosen to investigate the applicability of the resulting regression tree, for meteorological comparability of air pollutant concentrations by location (Fig. 1): Long Beach (N. Long Beach, 7 km North from Port of Long Beach and 25 km south from DTLA, 33.82°N/118.19°W); Pomona (middle of the SoCAB, 45 km east from DTLA, 34.07°N/117.75°W); Upland (foothills of the San Gabriel mountains, 55 km east from DTLA, 34.10°N/117.63°W); Rubidoux (an inland site, 78 km east from DTLA, 34.00°N/117.42°W); and San Bernardino (an inland site closer to mountains, 88 km east from DTLA, 34.11°N/117.27°W). Details about these measurement sites, which are operated by the SCAQMD, can be found in the California Air Resources Board's (CARB) air monitoring network description (CARB, 2011).

Fig. 1. Map of the study area and locations of pollutant monitoring sites (white squares) and NCEP upper air meteorology data obtained (red circles). Map from Google Earth. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.3. Vehicular emissions

Atmospheric levels of traffic-related primary pollutants depend strongly on emission source strengths, which are a function of the vehicle fleet and its maintenance, as well as vehicle miles travelled and traffic patterns. Thus, if the modeling periods extend too long, results may be influenced by long-term changes in emission rates and/or the number of vehicles in the modeling area. Annual vehicle fuel consumption in the SoCAB gradually increased with time prior to 2005, but during 2005–2008, fuel consumption reached a plateau (CALTRANS, 2009; Fig. S2). In addition, the number of registered vehicles in the SoCAB has remained nearly constant after 2007, decreasing minimally from 13,495,744 in 2007 to 13,278,657 in 2010 (DMV, 2011). Thus, it is expected that vehicle fuel consumption and the emission source strength did not change significantly from 2007 to 2009, the period examined here. Indeed, the Kolmogorov–Smirnov (KS) test showed the annual distributions of both daily mean and maximum NO_x at the DTLA monitoring site were statistically identical during the 2007–2009 study period ($p \gg 0.05$) (Supplementary material S3). In addition, Zhu et al. (2004) reported no seasonal variations in traffic flows, or in the ratio of vehicle types (heavy-duty diesel vs. gasoline) on the I-710 and I-405 freeways (major north-south roadways in the western SoCAB, Fig. 1). Moreover, the annual diurnal traffic patterns for the I-10 (which runs east-west over the length of the SoCAB) and I-15 (north–south in the eastern SoCAB, Fig. 1), show only small monthly variations (<5% and <13%, respectively, Fig. S4 in supplementary material). Thus, assuming traffic patterns on these freeways are reasonably representative, it appears seasonal changes in emissions were modest over the study period. We also note our analysis further assumes stationary source emissions of oxides of nitrogen (NO_x) varied little over the relatively short study period of three years. Stationary sources contribute less than 5% of CO emissions in the SoCAB and hence any changes in such emissions can be ignored (CARB, 2009).

In contrast to the lack of variation in seasonal and annual mean emission rates from traffic sources as discussed above, significant diurnal variations in vehicular emissions clearly occur. The annual mean diurnal profiles of traffic flow rates on the I-405 freeway show consistent patterns through the entire year (Fig. S4). Traffic flows reach a minimum around 03:00–04:00 and sharply increase with the onset of morning rush hours (04:30–07:00). This is followed by somewhat lower midday flows and a broad second peak in the late afternoon. Remarkably consistent diurnal patterns (scaled by total volume) have been observed for both the I-10 and I-15 freeways as well as several other freeways, indicating these are general traffic patterns throughout the majority of the SoCAB.

There are however significant differences in travel patterns and traffic flows between weekdays and weekends. To avoid day-of-the-week effects in vehicular emissions, only Tuesday–Friday data were collected and analyzed in this study. Mondays were also excluded to avoid possible carry-over effects from the previous weekend, as well as various Monday holidays.

Because emissions were not used as an explanatory variable in this analysis, the resulting regression trees cannot predict absolute

concentrations for days or locations involving different emissions patterns (e.g., weekend/holidays and other years with significant changes in emissions). Nonetheless, because a regression tree is created solely with meteorological variables that are not influenced by human activities such as emission changes, we can apply regression tree results to investigate meteorological comparability for years not in the study period as well as for days with different emissions (weekends/holidays).

We note that the design of the built environment can impact local pollutant concentrations, especially those of primary emissions. The monitoring stations we use data from, however, were sited to avoid direct influence from point or line sources (such as freeways). As a result one would not expect the stations to respond differently to particular weather patterns (i.e. one station has a high CO concentration when winds come from the north while others do not). Our analysis (below) confirms that the monitoring stations are not unduly influenced by local sources or building morphology (e.g., Fig. 6).

2.4. Primary pollutants as target variables

Of numerous pollutants emitted primarily from vehicular sources, the only species that are widely monitored are NO, NO_2 and CO. CO undergoes little reaction on time scales of hours, and for the purpose of this study is considered a conservative pollutant. Although NO is a highly reactive species, particularly during daytime when ozone concentrations are elevated, NO_x ($NO + NO_2$) is more conservative and a good indicator of vehicular emissions and atmospheric mixing. Thus, daily max. CO and NO_x concentrations ($[CO]_{max}$ and $[NO_x]_{max}$), and daily mean NO_x concentrations ($[NO_x]_{mean}$) at the DTLA monitoring site were chosen as representative target pollutants emitted from traffic sources. Due to the relatively low resolution for CO data, daily mean CO concentration was not considered in this analysis.

Although in the SoCAB nighttime traffic flows are only about 10% of daytime flows (Fig. S4), meteorological conditions, such as a stably stratified boundary layer and calm winds, allow pollutants to accumulate within the nocturnal boundary layer, resulting in higher concentrations of traffic-related pollutants such as NO_x and CO. The leading edge of the morning rush hour also contributes to pollutant concentrations in the early morning hours (Choi et al., 2012; Hu et al., 2009). Consistent with this, frequency histograms of $[CO]_{max}$ and $[NO_x]_{max}$ clearly show the maxima between 05:00 and 07:00 (Fig. S5). In the SoCAB, CO and NO_x concentrations also show strong seasonal variations, peaking in the winter and reaching a minimum in summer (Fig. 2a, b). This is primarily due to lower boundary layer heights and lighter wind speeds in the winter (Fujitani, 1986), particularly during the morning rush hour emission period when the sun rises later in winter and thus delays the onset of thermally induced mixing.

2.5. Meteorological variables as predictor variables

Most previous studies attempting to explain ozone or $PM_{2.5}$ concentrations with meteorological variables using statistical modeling methods found that fewer than 10 meteorological variables were significant predictors, among several tens of variables considered (Thompson et al., 2001). In the present study, a total of 29 upper-air and surface meteorological variables were used as inputs (Table 1), as follows: geopotential height (Φ) represents the synoptic-scale weather pattern; temperature at 850 mbar is a measure of the strength and height of the subsidence inversion; temperature differences between layers provide information about atmospheric stability; and the geopotential height gradients ($\Delta\Phi_{N-S} = \Phi_{north} - \Phi_{south}$ and $\Delta\Phi_{W-E} = \Phi_{west} - \Phi_{east}$) at 1000 mbar are likely to be strongly related to regional wind fields, and hence ventilation effects (Dye et al., 2003; Stoeckenius and Hudischewskyj, 1990). Air stability is likely related to surface temperature indirectly for nocturnal temperature inversions as well as for thermals in the convective boundary layer. Wind speeds are a measure of dispersion and ventilation strength and can affect boundary layer heights indirectly through turbulence intensity. Besides these parameters, relative humidity and surface pressure were added in the analyses as indirect meteorological factors. Some variables listed in Table 1 were additionally divided into daily, morning, and afternoon mean values to investigate intra-day effects.

The same set of meteorological data were used for all analyses performed here (Table 1). Upper air meteorological variables were extracted from the "4-times daily" National Centers for Environmental Prediction (NCEP) reanalysis database (Kalnay et al., 1996). Four data points (32.5°/35°N latitude and 117.5°/120°W longitude) around the SoCAB were selected and averaged to represent the upper air meteorological conditions above the SoCAB (Fig. 1). Surface weather variables for Los Angeles International Airport (LAX) were obtained from the MesoWest website operated by the University of Utah (http://mesowest.utah.edu/index.html). Fig. 2c shows a time-series of geopotential height at 925 mbar (Φ_{925mb}), and north–south pressure gradient at the 1000 mbar pressure level ($\Delta\Phi_{N-S}$). Also plotted are the surface meteorological variables, including daily mean wind speed (U_{mean}) and daily minimum temperature (T_{min}) with daily mean $[NO]_{mean}$ at Downtown LA as a reference in Fig. 2d.

Both upper-air and surface meteorological variables show strong seasonal variations, similar to $[CO]_{max}$ and $[NO_x]_{mean}$ (Fig. 2). Of the 18 upper air meteorological variables collected, pressure gradients ($\Delta\Phi_{N-S}$) and geopotential heights at 925 and 1000 mbar ($\Phi_{925mbar}$ and $\Phi_{1000mbar}$) show the best correlation with daily maximum $[NO_x]_{max}$, $[CO]_{max}$, and daily mean $[NO_x]_{mean}$ at the Downtown LA monitoring site, with correlation coefficients (r) ranging from 0.42 to 0.63. RH at 1000 mbar shows a significant negative correlation (−0.38 to −0.41) although its effect is indirect. Correlations with upper air temperature and wind speeds are less significant compared to other variables, with absolute correlation coefficients below 0.2.

Of the surface meteorological variables obtained at LAX, daily mean wind speed (U_{mean}), minimum temperature (T_{min}), and surface RH show strong negative correlations with NO_x and CO, with correlation coefficients of −0.51 to −0.60, −0.35 to −0.36, and −0.34 to −0.41, respectively. Daily mean surface pressure shows significant positive correlations ($r = 0.41$–0.44). Daily mean and daytime temperature effects are insignificant.

3. Regression trees results

3.1. Regression trees for the entire year

Once rain days, weekends, and Mondays were excluded, the number of 2007–2009 data points (days) input into the CART model were 553 and 549 for CO and NO_x, respectively. As noted, the regression trees explicitly show the effects of a specific meteorological parameter on pollutant levels. The CART analysis divided daily $[CO]_{max}$ into two subgroups based on the surface mean wind speed (U_{mean}) at the first split level, followed by geopotential height at 925 mbar ($\Phi_{925mbar}$), north-south Φ gradient ($\Delta\Phi_{N-S}$), daily minimum temperature (T_{min}), relative humidity at 1000 mbar ($RH_{1000mbar}$), and stability ($S_{925mbar}$), to make 11 final nodes (Fig. 3). For example, low U_{mean} generates less mechanical turbulence resulting in higher $[CO]_{max}$. Higher $\Phi_{925mbar}$ is related to the winter season (Fig. 2c), during which primary pollutant levels are typically elevated because of lower boundary layer

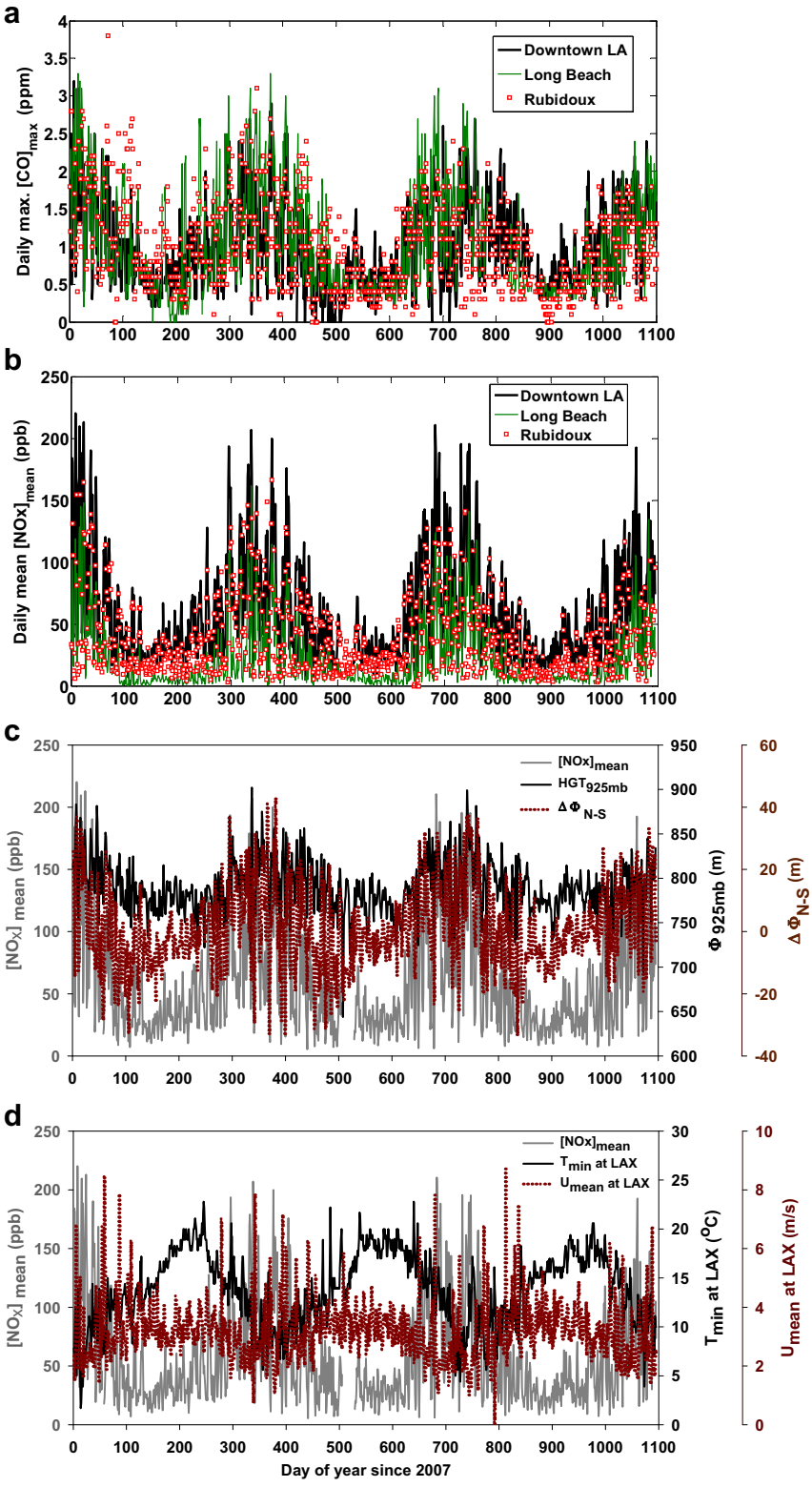

Fig. 2. Time-series of (a) daily maximum CO ($[CO]_{max}$), (b) daily mean NO_x concentrations ($[NO_x]_{mean}$) at downtown LA (black line), N. Long Beach (green line), and Rubidoux (red squares), (c) geopotential height at 925 mbar pressure level over the SoCAB (black solid line) and north–south geopotential height gradient at 1000 mbar (brown dotted line), and (d) surface daily minimum temperature (black solid line) and daily mean wind speed (brown dotted line) observed at Los Angeles International Airport (LAX). Light gray solid lines in (c) and (d) represent daily mean $[NO_x]_{mean}$ at downtown LA (N. Main St.) for a comparison. *x*-Axis is day of year since 2007 (Jan. 1, 2007 equals to 1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Meteorological variables used as explanatory (predictor) variables in the CART model and their effects on atmospheric primary pollutant concentrations.

Meteorological variables		Importance on primary pollutant level
Upper-air (NCEP model)	• Geopotential heights (Φ) at 1000/925/850/500 mbar	Indicator of synoptic-scale weather pattern
	• Mean temperature (T) at 1000/925/850 mbar	A measure of the strength and height of the subsidence inversion
	• Stability ($T_{1000\text{mbar}} - T_{925\text{mbar}}$, $T_{1000\text{mbar}} - T_{850\text{mbar}}$)	Indicator of atmospheric stability
	• Thickness ($\Phi_{925\text{mbar}} - \Phi_{1000\text{mbar}}$)	Related to the mean temperature in the layer
	• Relative humidity at 1000 mbar ($RH_{1000\text{mbar}}$)	Indirect effect
	• Pressure gradient at 1000 mbar level ($\Phi_{\text{north}} - \Phi_{\text{south}}$, $\Phi_{\text{east}} - \Phi_{\text{west}}$)	Related to wind fields and ventilation strength
Surface observations (LAX)	• Mean/min./max. temperature (T_{mean}, T_{min}, T_{max})	Indirect effects on air stability and emission rates from the engine
	• Mean/max. wind speed (U_{mean}, U_{max})	Related to dispersion/ventilation strength
	• Relative humidity (RH)	Indirect effect
	• Mean surface pressure	Indicator of synoptic-scale weather

heights, weaker winds, and possibly less active chemical sinks. Although surface layer temperature is not a direct function of atmospheric stability, surface temperature can be representative of surface cooling or heating. Enhanced surface heating can produce a deeper boundary layer and stronger turbulent energy during daytime and enhanced surface cooling can affect nocturnal atmospheric stability, showing an inverse correlation with pollutant concentrations.

$S_{925\text{mbar}}$ is defined as the temperature difference between 1000 mbar and 925 mbar pressures. A larger positive value of $S_{925\text{mbar}}$ represents less stable air due to warmer air below, implying enhanced mixing, and modest $[CO]_{max}$ concentrations. Interestingly, strong $\Delta\Phi_{N-S}$ also appears to be closely related to lower $[CO]_{max}$ whereas the west–east 1000 mbar geopotential height gradient $\Delta\Phi_{W-E}$, showed a positive correlation with $[CO]_{max}$. Steeper pressure gradients generally correlate with strong winds. However, wind fields in the SoCAB are dominated by a sea-breeze wind system oriented west–east. Thus, it is likely that prevailing westerlies or easterlies may be dampened by a strong north-south pressure gradient, establishing calm weather conditions with elevated primary pollutant concentrations. Otherwise, $\Delta\Phi_{N-S}$ may represent synoptic weather patterns related to calm meteorological conditions in the SoCAB. The regression tree for $[CO]_{max}$ at DTLA reproduces the observations well; the correlation coefficient between observations and representative nodal average values is 0.79. The mean absolute error is estimated to be 0.28 ppm, which is equivalent to of standard error of 28%.

Regression trees for $[NO_x]_{max}$ and $[NO_x]_{mean}$ were also created (Fig. 4 and S6). The first two splits for $[NO_x]_{max}$ are based on U_{mean} (1st split) and $\Phi_{925\text{mbar}}$ and $\Delta\Phi_{N-S}$ (2nd splits) exactly matching the initial split for $[CO]_{max}$. For the higher wind speed regime ($U_{mean} > 2.64$ m s^{-1}) of the first split, the final nodes are almost identical to those of the $[CO]_{max}$ regression tree (Fig. 4). Subsequent splits for $[NO_x]_{max}$ in the lower wind speed regime ($U_{mean} < 2.64$ m s^{-1}) are slightly different from those of $[CO]_{max}$. In this regime, $[NO_x]_{max}$ was divided by wind speed at the 3rd level, followed by RH_{LAX}, $\Delta\Phi_{W-E}$, and T_{min}. We also note that T_{min} also splits $[CO]_{max}$ in the lower wind regime by almost same criterion,

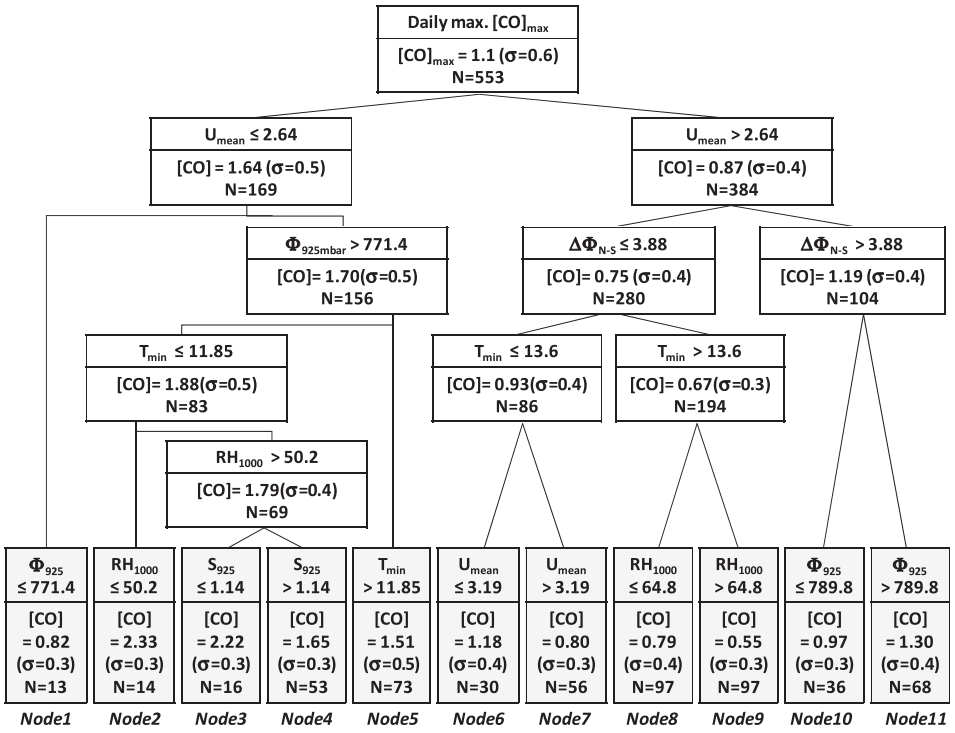

Fig. 3. Regression trees for daily $[CO]_{max}$ observed at downtown LA (N. Main St.) for 2007–2009. The split criteria of explanatory variables are shown at the top of each box (node). The bottom layer of each node indicates the mean $[CO]_{max}$ and standard deviation (σ) as well as the number of data in the node (N). Gray boxes represent the terminal nodes.

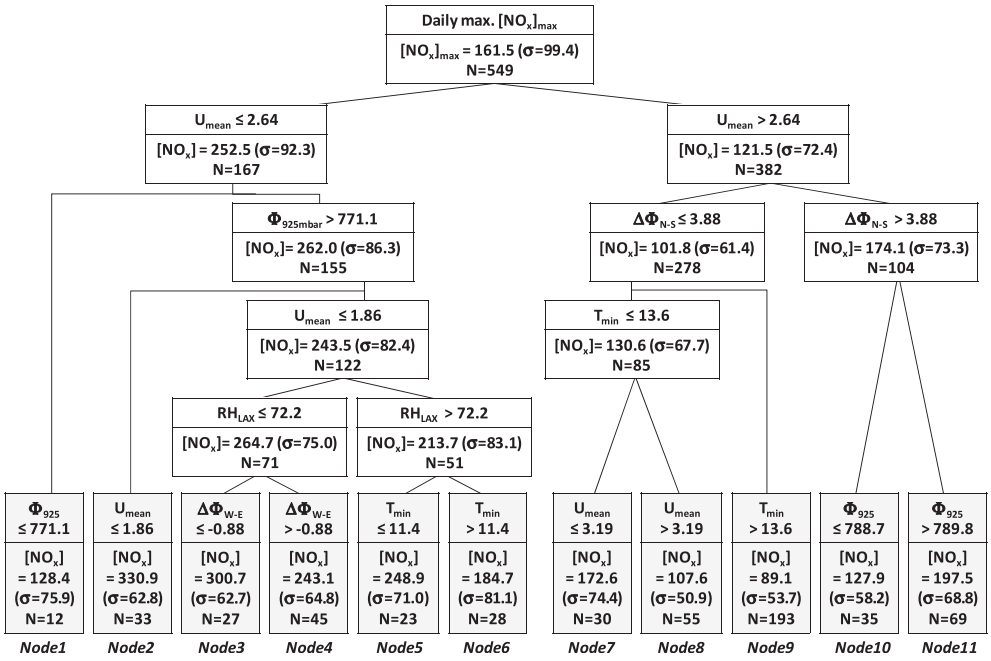

Fig. 4. Regression trees for daily $[NO_x]_{max}$ observed at downtown LA (N. Main St.) for 2007–2009. See Fig. 3 caption for an explanation of the notation.

although T_{min} is a more important variable for $[CO]_{max}$ (upper level split).

The regression tree for $[NO_x]_{max}$ has 11 final nodes, with almost identical major splits as the $[CO]_{max}$ regression tree. This similarity in $[NO_x]_{max}$ and $[CO]_{max}$ regression trees supports the validity of the CART model for traffic-related primary pollutants in urban areas. The correlation between observations and representative nodal average values ($r = 0.78$) for $[NO_x]_{max}$ is comparable to that for $[CO]_{max}$, although the mean absolute error and standard errors are slightly larger (48.7 ppb and 30%, respectively).

The daily $[NO_x]_{mean}$ regression tree houses 12 final nodes, first split by U_{mean} (2.64 m s^{-1}) and followed by U_{day}, $\Delta\Phi_{N-S}$ (2nd split), U_{mean}, Φ_{925mb} (3rd split), $RH_{1000mbar}$, $S_{925mbar}$, and RH_{LAX} where U_{day} is the daytime (10:00–16:00) mean surface wind speed. The correlation coefficient between actual and representative nodal average values is robust ($r = 0.87$) and the mean absolute error is estimated by the model to be 16.8 ppb (25% standard error) (Fig. S6). The $[NO_x]_{mean}$ regression tree also has several branches in common with the $[NO_x]_{max}$ regression tree.

It is axiomatic that many meteorological variables tend to be related to each other. Thus, it is not surprising that different pollutants have somewhat different meteorological variables in their optimized regression trees. For example, wind fields arise primarily from pressure gradients, and hence one pollutant tree may be slightly better divided by wind speed while the other is better divided by pressure gradient. In order to evaluate the comparability of the regression trees between the primary pollutants under consideration, mean $[NO_x]_{max}$ and $[NO_x]_{mean}$ were calculated for the days corresponding to each terminal node of the $[CO]_{max}$ regression tree. Excellent linear correlations between $[CO]_{max}$ and both $[NO_x]_{max}$ ($r = 0.99$) and $[NO]_{mean}$ ($r = 0.97$) (Fig. 5a, b) imply that the $[CO]_{max}$ regression tree can also

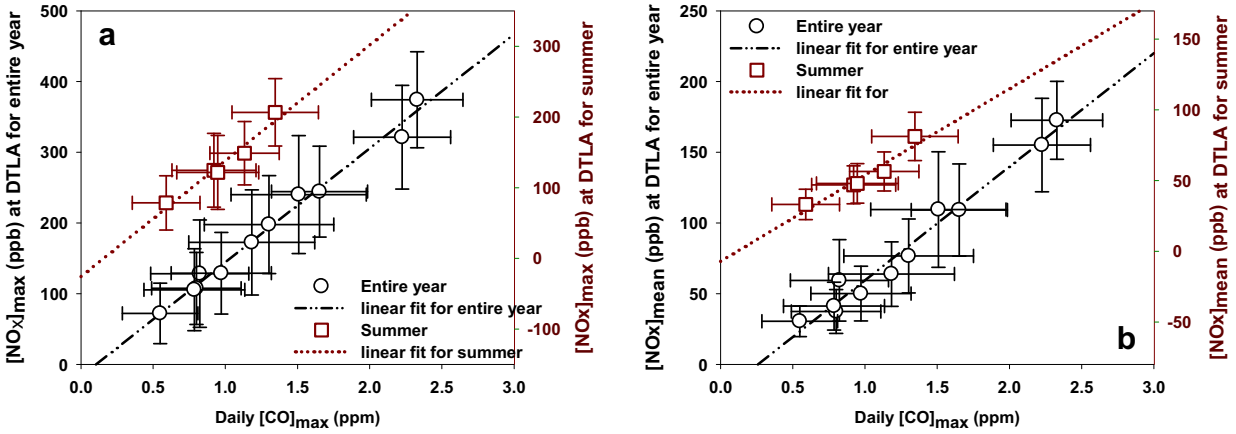

Fig. 5. Comparison of plots of the mean nodal $[CO]_{max}$ vs. (a) mean $[NO_x]_{max}$ and (b) $[NO_x]_{mean}$ for days that fall into the terminal nodes of the $[CO]_{max}$ regression tree at Downtown LA. Black circles represent the regression tree results for the entire year and dark red squares denote the seasonal regression trees for the summer (June 21–September 21). Horizontal and vertical bars denote standard deviation of $[CO]_{max}$ and $[NO_x]_{max}$ or $[NO_x]_{mean}$ in each terminal node, respectively. Black dash-dot line indicates linear fits for the entire year regression tree ($r = 0.99$ and 0.97 for $[NO_x]_{max}$ and $[NO_x]_{mean}$, respectively). The summer season regression tree yielded $r = 0.97$ and 0.92 for $[NO_x]_{max}$ and $[NO_x]_{mean}$, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

effectively split $[NO_x]_{max}$ and $[NO_x]_{mean}$, and that the $[NO_x]$ and $[CO]_{max}$ regression trees do classify meteorological conditions similarly.

3.2. Summer season regression trees

As noted earlier, concentrations of primary pollutants are lower and have lower variability in the summer season compared with other periods of the year primarily due to deeper boundary layer, as well as stronger thermally induced turbulence and higher surface wind speeds. Eighty-four percent of the total summer days fall into only three nodes (nodes 8, 9, and 10) of the full-year regression tree. To examine this season in detail and investigate if smaller standard deviations could be obtained with more focused regressions, summer season regression trees were created separately using the same explanatory variables as above (Figs. S7–S9 in supplementary material). Summer was defined as 21 June–21 September. Five final nodes were created for $[CO]_{max}$ and $[NO_x]_{max}$, and seven nodes for $[NO_x]_{mean}$.

Unlike the regression trees for the entire year, primary pollutant concentrations tend to be higher with higher surface temperature within the summer period. This inverse trend is likely due to the fact that higher temperature is generally linked to enhanced stagnation of air masses during the summer in the SoCAB (Stoeckenius and Hudischewskyj, 1990). Indeed, daily mean temperature was positively correlated with pollutant concentrations and negatively correlated with surface wind speeds for the summer season (Fig. S10 in supplementary material). The effects of other explanatory variables on concentrations for summer regression trees were similar to the entire year regression trees. Even for the summer, one predominant node appears, including about 60% of the summer days. Nonetheless, standard deviations in each final node were notably reduced for all pollutants for the summer regression trees (Fig. 5a, b).

4. Application to air quality studies

Because one of the advantages of the CART analysis is its effectiveness in explaining the variations in pollutant levels solely by a combination of meteorological conditions, regression trees can identify specific meteorological conditions that lead to low or elevated pollutant concentrations. As a result, the method can provide a sound basis for determining the meteorological comparability between measurement days. For example, if two days of a 6-day sampling period fall into the same final node of the regression tree and the other 4 days fall into a different final node, there are strong grounds for separating the entire measurement period into two meteorologically separate groups of days.

Investigating the meteorological comparability between different locations is more complicated due to different prevailing surface meteorology (e.g. the coastal area is generally influenced by stronger wind speeds and cooler surface temperature compared to non-mountainous inland areas in the SoCAB). Nonetheless, a mesoscale weather system governs surface meteorology across the region, and can cause similar day-to-day variability within the SoCAB (albeit with different absolute values). To investigate this further, we tested the robustness of the CART approach applied to spatially heterogeneous primary pollutants across several monitoring stations in the SoCAB. $[CO]_{max}$ data from each of the five sites in SoCAB were averaged for sets of days corresponding to the nodes of the regression tree developed using the regional meteorological data and $[CO]_{max}$ at downtown Los Angeles (Fig. 6). Overall, the $[CO]_{max}$ concentrations at DTLA were well correlated to average $[CO]_{max}$ values for all five monitoring sites, with correlation coefficients $r = 0.99$ for Long Beach; 0.97 for Pomona; 0.86 for Upland; 0.78 for Rubidoux; and 0.91 for San Bernardino. Slopes varied somewhat between locations. We note that correlation coefficients are significantly improved if the one notably deviated point, node 2, is excluded ($r = 0.99$ for Pomona; 0.97 for Upland; 0.91 for Rubidoux; 0.98 for San Bernardino). These results clearly show that high or low pollution episodes are generally coincident across the SoCAB, and support the applicability of the CART analysis to comparisons of meteorological conditions linked to variations in primary pollutant concentrations by location, provided the period of node 2 is excluded.

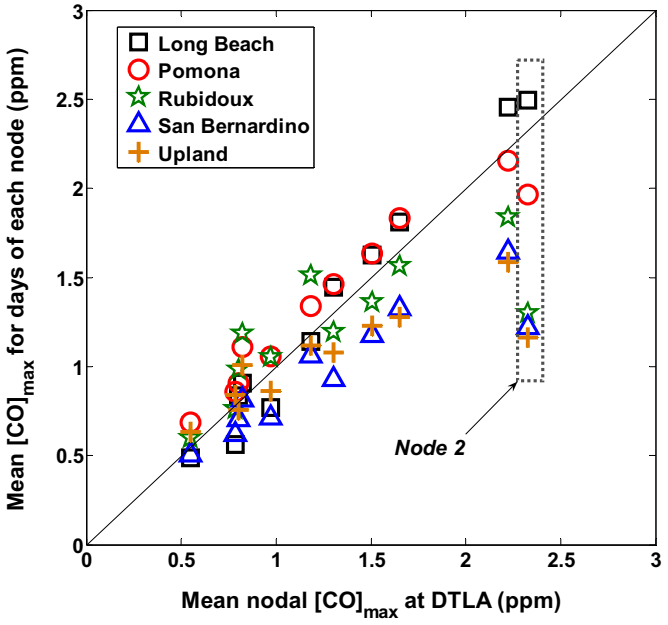

Fig. 6. 1:1 comparison plot between the mean nodal $[CO]_{max}$ at downtown LA (DTLA) vs. mean concentrations for days that fall into corresponding terminal nodes at five monitoring sites in the SoCAB (Long Beach, black squares; Pomona, red circles; Rubidoux, green stars; San Bernardino, blue triangles; and Upland, orange crosses). The gray solid line represents a 1:1 relationship. The dotted box represents *node 2* which showed a distinct deviation at the eastern sites (see text). The correlation coefficient (r) of 0.99, 0.97, 0.78, 0.91 and 0.86 were obtained between DTLA vs. Long Beach, Pomona, Rubidoux, San Bernardino, and Upland, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Deviations for node 2 were strongest for the eastern monitoring sites located closer to mountainous areas (Fig. 6). For this period, which consists of 3 spring days and 11 winter days, a relatively high average $[CO]_{max}$ was observed at DoLA (2.33 ppm) while $[CO]_{max}$ was relatively low at the eastern sites. These days were characterized by unusually high pressure both at the surface and at 1000 mbar, the lowest humidity of the winter season, and relatively high temperatures for the winter period (Table 2). Based on these meteorological peculiarities, node 2 generally represents the locally well-known weather pattern referred to as "Santa Ana winds." Santa Ana winds are largely orthogonal in direction to the more typical land breeze-sea-breeze winds, and result from the build-up of high pressure over the high altitude Great Basin between the Rocky and Sierra-Nevada mountains, which generates strong mountain down slope winds in the SoCAB. When the air is forced down slope, adiabatic heating results in increasing temperature and dropping relative humidity. At the same time, winds are intensified as the air is channeled through canyons (Fovell, 2002). Although Santa Ana's are characterized by strong winds, there can be considerable spatial and temporal variations in wind strength with strong flow in the vicinity of the mountains and canyons, but weaker flows as the air spreads out across the coastal plain (Sommers, 1978). Consequently, strong dispersion of pollutants

Table 2
Number of days and the mean meteorological values in the final nodes of the $[CO]_{max}$ regression trees developed at downtown LA (N. Main St.). Node 2 shows the highest geopotential heights, surface pressure, lowest relative humidity, and higher temperature during the winter season.

	$[CO]_{max}$	Number of days				Mean meteorological variables in the node					
		Spring (Mar–May)	Summer (Jun–Sep)	Fall (Oct–Nov)	Winter (Dec–Feb)	U_{mean}	T_{day}	RH_{LAX}	RH_{1000}	Φ_{1000}	P_{LAX}
Node 1	0.82	3	5	1	4	2.3	17.7	75.5	67.5	97.3	1012
Node 2	*2.33*	*3*	*0*	*0*	*11*	*1.9*	*20.5*	*40.9*	*45.3*	*174.6*	*1020*
Node 3	2.23	2	0	1	13	2.0	17.4	72.8	64.2	149.2	1017
Node 4	1.65	5	0	6	42	2.2	16.3	66.8	65.9	167.1	1019
Node 5	1.51	6	12	46	9	2.2	22.8	59.3	55.4	136.7	1015
Node 6	1.18	26	0	3	1	2.9	17.6	71.4	66.9	136.3	1016
Node 7	0.80	42	2	2	10	4.2	16.3	66.0	68.1	135.2	1016
Node 8	0.79	11	82	0	4	3.3	21.7	73.7	60.0	104.3	1012
Node 9	0.55	20	64	0	13	3.6	20.4	73.5	70.0	105.9	1013
Node 10	0.97	9	21	4	2	3.4	21.5	68.5	57.7	100.0	1011
Node 11	1.30	16	14	12	26	3.3	19.1	55.5	58.6	154.1	1017

results in low concentrations near the mountains (e.g. Rubidoux, San Bernardino, and Upland). However, the winds weaken as they spread out across the coastal plain, and during these "flow reversal" events coastal areas are frequently downwind of more of the urbanized areas of the basin, and thus may experience elevated primary pollutant concentrations. Therefore, the heterogeneity in the east–west spatial distributions of primary pollutants in SoCAB on an annual basis is enhanced by the Santa Ana wind system, which does not generally affect north-south spatial distributions as strongly (e.g., DTLA and Long Beach).

CART has some potential to provide a statistical basis for meteorological adjustment of primary pollutant data collected on individual days. This, in turn, may facilitate more robust comparisons between areas with sparse datasets. For example, assume we obtained $[CO]_{max} = A$ at SoCAB location i on day 1, which falls into node 8 in the $[CO]_{max}$ CART, and also measured $[CO]_{max} = B$, where $A < B$, at location j on day 2, which has meteorological characteristics typical of node 5. Based on the CART results, it is expected that $A < B$ due to purely meteorological effects, and thus depending on the magnitude of the difference between A and B, it may not be appropriate to say location j is more polluted merely because B is greater than A in the measurement data. If, however, we normalized the two datasets based on the ratio of the representative concentrations of the two nodes, equal to 1.51(node 5)/0.79(node 8) = 1.92, meteorologically adjusted concentration $1.92A$ can be more reasonably compared to B. The adjustment should be done with caution however since the ranges for each node are significant, and the slopes of the relationships between pollutants (e.g., Fig. 6) deviate somewhat from unity.

Meteorologically adjusted concentrations may also be useful for investigating long-term trends of pollutant levels. The CART model offers the capability to investigate the long-term trends in various primary pollutants grouped by meteorological classification, similar to meteorologically adjusted ozone studies carried out in the past (Thompson et al., 2001).

5. Conclusion

We have demonstrated that a CART approach for primary pollutant concentrations results in a useful tool to determine the comparability, or lack thereof, of meteorological conditions between locations and across days in the SoCAB. It is likely this approach can be applied to other air basins as a basis for integrating vehicle-related air pollutant measurements conducted at various locations and different times, however more work is needed to establish the ability of the CART approach to produce useful insights in areas with more variable weather compared to the SoCAB.

Acknowledgements

The authors gratefully acknowledge support for this study by the California Air Resources Board (gs1), Contract No. 09-357. This study was made possible in part due to the data made available by the governmental agencies, commercial firms, and educational institutions participating in MesoWest. Helpful comments of two anonymous reviewers greatly improved the manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.atmosenv.2012.09.049

References

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.G., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, California, USA.
CALTRANS, 2009. 2008 California Motor Vehicle Stock, Travel and Fuel Forecast, Appendix C Vehicle Fuel Consumption by County and Fuel Type. California Deartment of Transportpation, Sacramento.
CARB, 2009. 2008 Estimated Annual Average Emissions. California Air Resources Board, Sacramento, CA.
CARB, 2011. Quality Assurance Air Monitoring Site Information. CARB. http://www.arb.ca.gov/qaweb/siteinfo.php.
Choi, W., He, M., Barbesant, V., Kozawa, K.H., Mara, S., Winer, A.M., Paulson, S.E., 2012. Prevalence of wide area impacts downwind freeways under pre-sunrise stable atmospheric conditions. Atmospheric Environment 62, 318–327.
De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.
DMV, 2011. Estimated Fee-paid Vehicle Registrations by County, California. DMV Forecasting Unit, California Department of Motor Vehicles.
Dye, T.S., MacDonald, C.P., Anderson, C.B., Hafner, H.R., Wheeler, N.J.M., Chan, A.C., 2003. Guidelines for Developing an Air Quality (Ozone and PM2.5) Forecasting Program, EPA-456/R-03-002. Environmental Protection Agency, NC, USA.
Fovell, R., 2002. The Santa Ana Winds. Available from: http://www.atmos.ucla.edu/~fovell/ASother/mm5/SantaAna/winds.html.
Fujitani, T., 1986. Seasonal-variation of the structure of the atmospheric boundary-layers over a suburban area. Atmospheric Environment 20, 1867–1876.
Hess, K.R., Abbruzzese, M.C., Lenzi, R., Raber, M.N., Abbruzzese, J.L., 1999. Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma. Clinical Cancer Research 5, 3403–3410.
Hoek, G., Boogaard, H., Knol, A., De Hartog, J., Slottje, P., Ayres, J.G., Borm, P., Brunekreef, B., Donaldson, K., Forastiere, F., Holgate, S., Kreyling, W.G., Nemery, B., Pekkanen, J., Stone, V., Wichmann, H.E., Van der Sluijs, J., 2010. Concentration response functions for ultrafine particles and all-cause mortality and hospital admissions: results of a European expert panel elicitation. Environmental Science & Technology 44, 476–482.
Horie, Y., 1988. Air Quality Management Plan 1988 Revision, Appendix V-P: Ozone Episode Representativeness Study for the South Coast Air Basin, El Monte, CA.
Hu, S., Paulson, S.E., Fruin, S., Kozawa, K., Mara, S., Winer, A.M., 2012. Observation of elevated air pollutant concentrations in a residential neighborhood of Los Angeles California using a mobile platform. Atmospheric Environment 51, 311–319.

Hu, S.S., Fruin, S., Kozawa, K., Mara, S., Paulson, S.E., Winer, A.M., 2009. A wide area of air pollutant impact downwind of a freeway during pre-sunrise hours. Atmospheric Environment 43, 2541–2549.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., Joseph, D., 1996. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society 77, 437–471.

Kozawa, K.H., Fruin, S.A., Winer, A.M., 2009. Near-road air pollution impacts of goods movement in communities adjacent to the Ports of Los Angeles and Long Beach. Atmospheric Environment 43, 2960–2970.

Krudysz, M., Moore, K., Geller, M., Sioutas, C., Froines, J., 2009. Intra-community spatial variability of particulate matter size distributions in Southern California/Los Angeles. Atmospheric Chemistry and Physics 9, 1061–1075.

Moore, K., Krudysz, M., Pakbin, P., Hudda, N., Sioutas, C., 2009. Intra-community variability in total particle number concentrations in the San Pedro Harbor Area (Los Angeles, California). Aerosol Science and Technology 43, 587–603.

Morgan, J.N., Sonquist, J.A., 1963. Problems in an analysis of survey data and a proposal. Journal of the American Statistical Association 58, 415.

Penttinen, P., Timonen, K.L., Tiittanen, P., Mirme, A., Ruuskanen, J., Pekkanen, J., 2001. Ultrafine particles in urban air and respiratory health among adult asthmatics. European Respiratory Journal 17, 428–435.

Pope, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D., 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Jama-Journal of the American Medical Association 287, 1132–1141.

SCAQMD, 2012. 2012 Air Quality Management Plan (AQMP). South Coast Air Quality Management District, Diamond Bar, CA.

Sommers, W.T., 1978. LFM forecast variables related to Santa-Ana wind occurrences. Monthly Weather Review 106, 1307–1316.

Stoeckenius, T.E., Hudischewskyj, A.B., 1990. Adjustment of Ozone Trends for Meteorological Variation. U.S. Environmental Protection Agency, Research Triangle Park, NC.

Thompson, M.L., Reynolds, J., Cox, L.H., Guttorp, P., Sampson, P.D., 2001. A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmospheric Environment 35, 617–630.

Tooke, T.R., Coops, N.C., Goodwin, N.R., Voogt, J.A., 2009. Extracting urban vegetation characteristics using spectral mixture analysis and decision tree classifications. Remote Sensing of Environment 113, 398–407.

Turner, J.R., Allen, D.T., 2008. Transport of atmospheric fine particulate matter: Part 2-Findings from recent field programs on the intraurban variability in fine particulate matter. Journal of the Air & Waste Management Association 58, 196–215.

Venkatram, A., 2004. The role of meteorological inputs in estimating dispersion from surface releases. Atmospheric Environment 38, 2439–2446.

Venkatram, A., Cimorelli, A.J., 2007. On the role of nighttime meteorology in modeling dispersion of near surface emissions in urban areas. Atmospheric Environment 41, 692–704.

Zhang, K.M., Wexler, A.S., Zhu, Y.F., Hinds, W.C., Sioutas, C., 2004. Evolution of particle number distribution near roadways. Part II: the 'road-to-ambient' process. Atmospheric Environment 38, 6655–6665.

Zhu, Y.F., Hinds, W.C., Shen, S., Sioutas, C., 2004. Seasonal trends of concentration and size distribution of ultrafine particles near major highways in Los Angeles. Aerosol Science and Technology 38, 5–13.

Zhu, Y.F., Kuhn, T., Mayo, P., Hinds, W.C., 2006. Comparison of daytime and nighttime concentration profiles and size distributions of ultrafine particles near a major highway. Environmental Science & Technology 40 (8), 2531–2536.