On the Ratio of Sulfur Dioxide to Nitrogen Oxides as an Indicator of Air Pollution Sources

RONIT NIREL AND URI DAYAN

The Hebrew University of Jerusalem, Jerusalem, Israel

(Manuscript received 24 March 2000, in final form 15 November 2000)

ABSTRACT

The ratio of sulfur dioxide to nitrogen oxides (RSN = SO_2/NO_x) is one indicator of air pollution sources. The role of this ratio in source attribution is illustrated here for the Ashdod area, located in the southern coastal plain of Israel. The main sources of pollution in the area are the tall stacks of the Eshkol power plant, the stacks of oil refineries, and areal sources (stationary and mobile). The factors that affect RSN are studied using four regression models: a binary regression tree in original scale, a tree in logarithmic scale, a data partition produced by a combination of the two trees, and a linear regression model. All models have similar relative prediction error, with the combined partition best highlighting the sources of variability in RSN: (a) very low values (interquartile range of [0.12, 0.48]) are associated with traffic, (b) low values ([0.43, 1.00]) are attributed to the power plant and to daytime emissions of local industry, (c) medium values ([0.74, 1.90]) are associated with local industry emissions during cooler hours of the day and refinery emissions mainly on slow wind episodes, and (d) high values ([1.07, 4.30]) are attributed to refinery emissions during moderate to fast wind episodes. Analysis of the number of episodes of increased concentrations indicates that, during 1996 and 1997, about 42% of SO_2 episodes are attributable to the power plant and 33% to the refineries. Increased-NO_x episodes are mainly contributed by traffic (91%) and power plant (4.5%) emissions.

1. Introduction

In many regions, air pollution is caused by different types of sources. One type is point sources, such as tall stacks of power plants and oil refineries stacks. A second type is areal sources, such as local industry and transportation. The ability to differentiate among various sources can be useful (a) for quantifying the relative contribution of each source, in particular the relative contribution of mobile sources; (b) for evaluating the effectiveness of existing control measures; and (c) for gaining insight into the mechanisms and conditions that affect the level of pollution emitted by different sources.

The level, location, and duration of pollutants' concentration within a region depend on plume height, wind speed, rate of vertical mixing in the atmosphere, and distance from the source. Hence, plumes from different sources can be identified by the following indicators:

- *Structure of the atmospheric boundary layer (ABL).* During the day, the concentration of pollutants emitted from areal sources is well mixed throughout the boundary layer. This mixing is more vigorous and efficient within the free convective layer (FCL). In this layer, free convection is formed when buoyancy dominates turbulence production rather than shear. Such convective heating is manifested by a pronounced superadiabatic lapse rate of up to $-3°$ to $-3.5°C\ (100\ m)^{-1}$, as measured by Dayan et al. (1988) in the southern part of the coast of Israel. Plumes injected above the ABL from tall point sources may drift downward and cause localized high concentrations by impingement of the plume upon the ground as a result of convective downdrafts. As opposed to the FCL, which is driven by thermal turbulence, a stable boundary layer (SBL) may be built during the night in clear sky conditions due to radiative heat losses from the ground and cooling by conduction. Dayan and Koch (1989) found that the frequency of radiative inversions along the coast of Israel for each of the three synoptic categories occurring during the spring and beginning of summer were in the range of 13%–22%, whereas the frequency for each of the other synoptic categories was less than 7%. Such a stable boundary layer prevents plumes emitted from tall point sources from reaching the surface and therefore does not affect the surface air quality, while emissions from mobile and other ground sources are trapped within the stable layer and increase the level of pollution.
- *Wind direction.* Frequently, pollution sources can be identified by the upwind direction from the monitoring

Corresponding author address: Dr. Ronit Nirel, Dept. of Statistics, The Hebrew University of Jerusalem, Mt. Scopus, Jerusalem 91905, Israel.
E-mail: nirelr@cc.huji.ac.il

© 2001 American Meteorological Society

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

station, or receptor. In extracting the actual wind direction from surface measurements, the height of the source, its distance from the receptor, and the measurement timescale should be taken into account. Above the coastal plain of Israel, typically observed veering angles in wind profiles within the ABL (i.e., from the surface up to the frictionless layer) are 5°–10° for neutral atmospheric conditions and 30°–40° for slightly stable conditions (Balmor et al. 1988; Koch and Dayan 1992). Besides veering of the horizontal direction of the wind profile, two other parameters should be considered: the deviation σ_θ of the horizontal wind direction fluctuation at the surface and the standard deviation σ_y of the concentration distribution in lateral direction as a function of the distance from the pollution sources.

- *Plume signature: RSN = SO_2/NO_x.* Gas-phase sulfur dioxide (SO_2) is emitted during combustion of all sulfur-containing fuels (oil, coal, and diesel), whereas traffic is a prominent source of nitrogen oxides (NO_x). The ratio between the two may be useful in identifying pollution sources, because fuels used, say, for electricity generation and for transportation differ in their sulfur content and because the ratio is related to combustion conditions. Typically, electricity production is expected to result in a lower SO_2/NO_x ratio than emissions caused by low-temperature boilers burning fuel oil with high sulfur content. Benkovitz et al. (1996) mapped the nitrogen-to-sulfur ratios (moles N emitted/moles S emitted) to reflect the overall character of pollution sources around the world. A study of the inverse ratio S/N indicates that, in many European countries, the ratio is typically near 1. It is larger than 3 in some countries in central and eastern Europe, reflecting, most likely, the presence of heavy industry. In industrialized and heavily populated areas of North America, the ratio is generally greater than 1, while in areas located away from large stationary sources, it is around 0.1. Studies in Georgia report a mean SO_2/NO_x ratio of 0.05 for mobile sources and ratios ranging from 2.7 to 4.6 for power plant plumes (Duncan et al. 1995). These findings indicate that mobile and point sources may be identified by their characteristic RSN.

Meteorological conditions cannot always help in identifying the sources that account for specific pollution episodes. For such cases, a source apportionment approach using receptor models is often used. Models of this kind may be of a chemical orientation, such as the chemical mass balance model, which provides a quantitative assessment of the contribution of emission sources to pollutant concentrations in the receptor (Watson et al. 1990). Another type of receptor model is based on statistical methods such as multiple linear regression, principle component analysis, or cluster analysis models (e.g., Swietlicki et al. 1996). Such models assume that, although ambient concentrations result from the superposition of several sources (Hopke 1985), a dominant source often can be identified.

In this paper, the factors that affect the variability of RSN are studied through regression tree models. The results are then used to classify pollution episodes by their probable source. Tree models are useful exploratory tools for generating insight into the nature of a relationship between response and explanatory variables. In comparison with linear models, tree-based models are more adept at capturing nonlinear behavior and allow more general interactions among predictors. We demonstrate the method for the Ashdod area, Israel. The different sources of pollution in the area and the data from an air quality monitoring station within this airshed are described in section 2. Section 3 describes regression tree models and compares them with linear regression. Four regression models are fitted and validated in section 4: a regression tree model in original scale, a tree model in logarithmic scale, a data partition produced by a combination of the two trees, and a linear regression model in log scale. The combined partition defines 20 data subsets, representing four RSN levels: very low (median RSN 0.21), low (0.62), medium (1.29), and high (2.11). The meteorological conditions underlying these levels are analyzed in section 5a. In section 5b, ''exceedances'' of SO_2 and NO_x concentrations beyond one-quarter of the respective air quality standards are classified by their probable pollution source. Section 6 summarizes the main findings.

2. Location, data, and descriptive analysis

We analyze data from the Ashdod air quality monitoring station, located in the southern coastal plain of Israel (Fig. 1). The main pollution sources in the area, within approximately a 20-km range of the receptor, are (a) point sources, including the 150-m stacks of the Eshkol power plant, located about 3 km (1.9 mi) northwest of the receptor, and high stacks (80 m) of the Ashdod oil refineries, located approximately 2.5 km to the north of the monitoring station; (b) mobile sources, including traffic on the Tel Aviv–Ashqelon highway to the east of the receptor, on the road to the Ashdod harbor north of the receptor, and on the road leading to the town west of the station; and (c) areal sources, including mainly small food and pharmaceutical industries. These sources are located mostly to the west of the receptor.

Identifying the pollution sources in this area may prove to be simple under some conditions; for example, during the hot hours of the day in the summer when pollution arriving from the northwesterly sector is attributable to the power plant. However, it is more difficult to distinguish between sources in the cooler hours of the day. At these times, the pollution intercepted by the station from the northern sector, for example, can be attributed to either stationary (refineries) or mobile (traffic) sources.

The data are composed of 30-min averages of SO_2

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

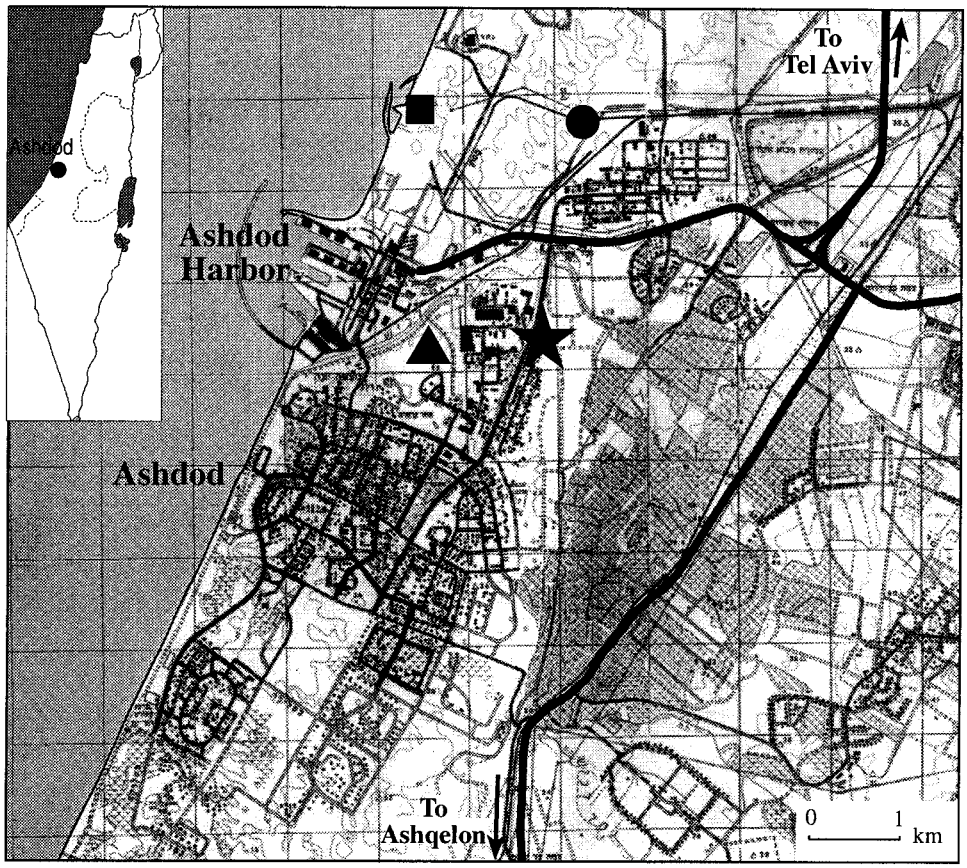

FIG. 1. The Ashdod–Hevel Yavne area, depicting the Ashdod air quality monitoring station (star), the power plant (square), the oil refineries (circle), and local industries (triangle).

and NO_x concentrations, as recorded in the period of 1987–97. Surface meteorological measurements include wind direction, wind speed, and temperature. Monthly fuel data specify the type of fuel (regular, low, and very low sulfur content) and quantities used by the power plant and refineries. Since 1996, the power plant fuel policy requires that fuel containing ''low'' sulfur content (1%) is to be used between April and October, and fuel containing ''regular'' sulfur content (2%) is to be used during the rest of the year. ''Very low'' sulfur content fuel (0.5%) should be used when the Israel Meteorological Service Intermittent Control System issues a pollution alert.

Between 1987 and 1997, the total fuel consumption of the power plant was fairly stable, but the percentage of low-sulfur fuels consumed increased continuously. This fact resulted in a pronounced decrease in sulfur emissions during this period and in a respective decline in RSN values intercepted downwind of the power plant (Fig. 2). The largest median value of RSN was observed in 1988 (about 2.8), when 50% of the values ranged approximately from 2 to 4. Since 1988, there was a continuous decline in the median RSN, and in 1996 and 1997 the levels were near 1.

In view of the high variability of RSN in the power plant wind sector, it seems reasonable to narrow the analysis to 1996 and 1997, which share the same fuel consumption policy. The values of RSN in 1996–97 range from 0 to 36.7. To eliminate the influence of negligible concentrations, in the numerator and denominator of RSN, only measurements with SO_2 greater than 20 $\mu g\ m^{-3}$ (this is background concentration in the area) and $0 < RSN < 15$ were included in the analysis (number of measurements $n = 5868$). Twenty-five episodes, clearly attributable to the refineries, had $RSN \geq 15$.

The diurnal, seasonal, and directional variation of RSN is summarized by the RSN roses in Fig. 3. Note that RSN values appear in a square root scale for a clearer exposition. Figure 3a illustrates mean RSN values under well-mixed FCL conditions (1100–1500 LST). During the hot hours of the day, pollution is intercepted from the western and northern sectors: in January, mean RSN values range from 0.39 to 1.40, in the 205°–45° wind sector; in April, the means range from 0.76 to 1.82, in the 215°–355° sector; in July, RSN means are between 0.28 and 1.33 in the 225°–335° sector; and in October, the means are in the range of 0.43–

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

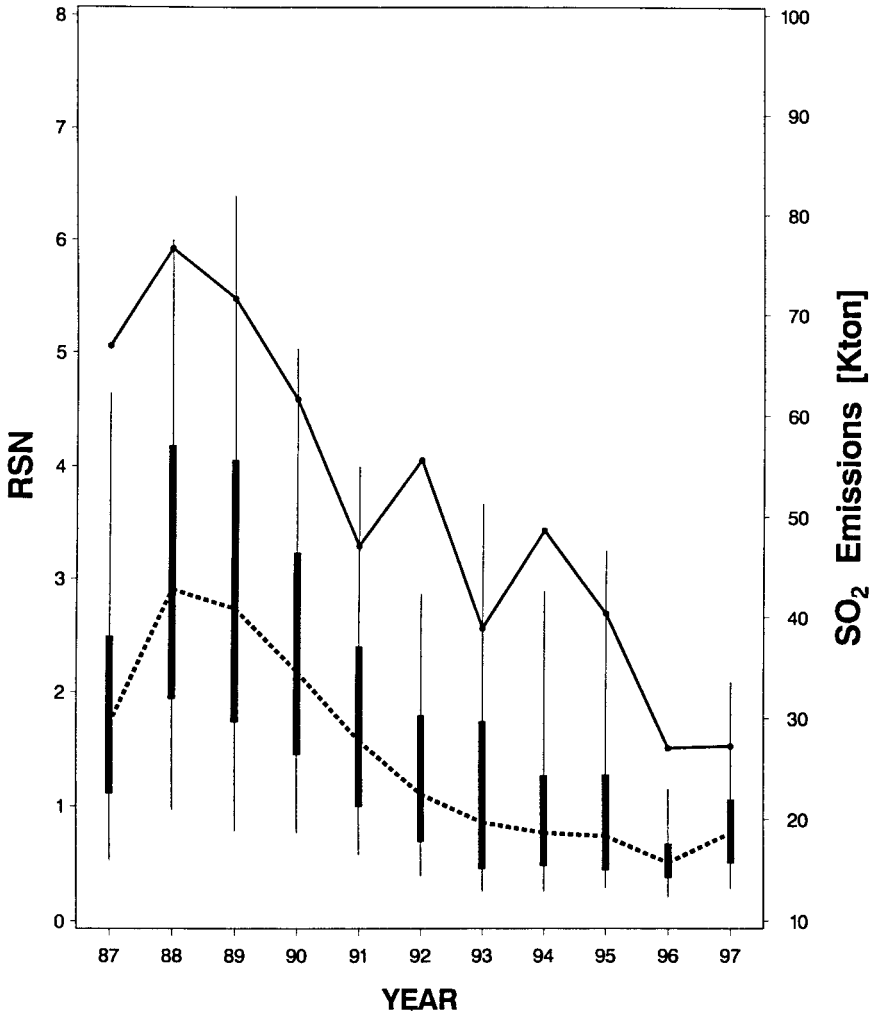

FIG. 2. Distribution of RSN attributable to the power plant for 1987–97. Included events satisfy: $SO_2 > 20$ μg m^{-3}, $0 <$ RSN < 15, wind sector within 280°–340°, and time span 0900–1800 LST on days for which the minimum temperature between 1100 and 1700 LST exceeds 23°C. The dashed line connects the median RSN values, and the solid line depicts annual amounts of SO_2 (metric kilotons) emitted by the power plant. The box indicates the interquartile range; the vertical lines extend to the 5th and 95th percentiles ($n = 6822$).

3.03 in the 225°–345° direction. Hence, daytime RSN levels are higher and more variable in the transitional seasons than in the winter and summer. Wintertime episodes are contributed by sources to the west (power plant and local industry) and north (refineries) of the receptor; during the rest of the year, pollution episodes are mainly attributable to westerly sources. Figure 3b illustrates the mean RSN values under a stratified SBL (1800–2200 LST). In January, pollution is intercepted from two wind sectors: within the 335°–55° sector, RSN means range from 0.24 to 2.02, while within the 225°–285° sector, they range from 0.48 to 0.60. In April, three pollution sectors are identifiable: the 315°–55° sector contributes RSN levels between 0.88 and 4.42; the mean levels in the 95°–135° sector are in the range of 0.15–0.29; and in the 245°–295° sector, 0.75–1.74. In July, pollution episodes are intercepted in the 255°–355° sector, with mean RSN in the range of 0.49–1.81. October episodes are in the 345°–65° sector, with RSN means ranging from 0.51 to 4.32. Thus, evening episodes in the study area have a higher spatial dispersion as compared with daytime episodes and have more variable RSN levels. The highest variability is observed in springtime, with high RSN values intercepted from the north and low values from the east (most likely attributable to transportation). The lowest variability is observed in the summer, when relatively stable RSN values are intercepted from the northwesterly sector. The observed spatial homogeneity in the summer is explained meteorologically by the fact that the summer in Israel is characterized by the predominant barometric trough known as the ''Persian trough.'' This unique synoptic system generates northwesterly winds almost every day along the coastal region of Israel.

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

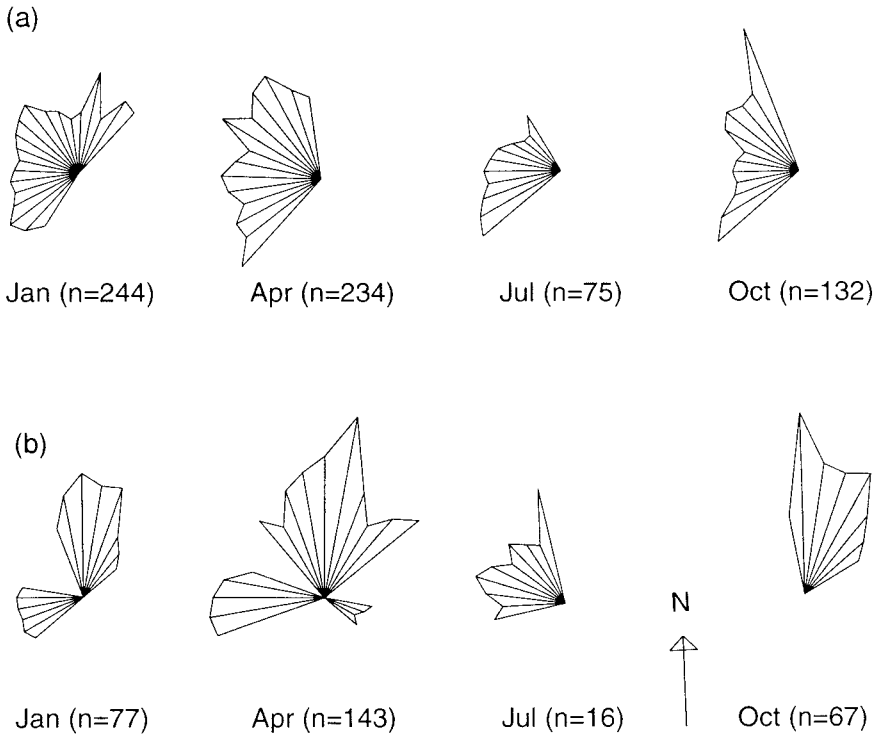

FIG. 3. Mean RSN by wind sector and season: (a) 1100–1500 LST and (b) 1800–2200 LST. Values are presented in a square root scale. The length of the arrow corresponds to RSN = 1.

3. Regression tree models

The sources of variability in the SO_2/NO_x ratios are investigated in this paper through regression tree models and the data partitions that they induce. The regression tree technique was developed by Breiman et al. (1984) in a wider framework of the Classification and Regression Trees (CART) method. Regression trees, like linear regression, study relationships between a response variable, or predictand, Y and a set of possible p predictors $X_1, \ldots, X_p$. The rules that define the relationship between the predictand and predictors are displayed in the form of a binary tree, hence the name. Tree-based models are descriptive exploratory tools, whereas linear regression models can be used for statistical inference, provided the distributional assumptions are satisfied. CART does not assume linearity and can therefore capture a wider variety of structures than the linear model. Questions of interaction between predictor variables are handled automatically by the tree-building process. In addition, for some applications, the tree structure gives results that are easier to interpret than the classical regression equation. These properties of tree-based models have made them attractive and useful for modeling environmental and meteorological data. Recent applications include Burrows et al. (1995) and Ryan (1995), who utilized CART models for prediction of ozone concentrations in southern Canada and Baltimore, respectively. Burrows (1997) built CART models for predicting UV radiation, Carter and Elsner (1997) forecast rainfall over Puerto Rico, and Walmsley et al. (1999) used CART to investigate the factors that affect liquid water content in fog.

The CART builds a binary decision tree by partitioning the data into disjoint subsets, or ''bins,'' represented by *nodes* in the tree. The first node of the tree, or the *root* node, is the entire dataset. The tree is built by splitting the root node into two new nodes. These nodes may be also split, and so forth. A node that is not split is referred to as a *terminal node.* This building process is often called recursive partitioning. Each successive partition in the tree is defined by a predictor X_i and a split value s, assigning observations for which $x_i \leq s$ to the left child node and those for which $x_i > s$ to the right (uppercase letters are used for random variables and lowercase for the observed values). At each partition, the algorithm examines every allowable split value on each of the predictors. The split that minimizes the prediction error is selected. The prediction for all observations in a terminal node i is equal to the predictand mean at this node.

The tree is grown until some stopping rule is satisfied (e.g., minimal size of a terminal node or error threshold). Frequently, the trees tend to grow too big with too few observations in each terminal node. To overcome this problem, a ''bottom-up'' process is initiated, which incrementally collapses back pairs of terminal nodes and is named recursive pruning. The final tree minimizes the cost complexity measure, which contains a penalty for tree size.

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

CART is an exploratory technique, and its main tool to assess model fit is the prediction error. Define the deviance in node i by

$$D_i^2 = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2,$$

where n_i is the number of observations in node i, y_{ij} the predictand value for the jth observation in node i, and $\bar{y}_i$ the observed mean of the predictand in node i. The overall deviance is given by

$$D^2 = \sum_{i=1}^{b} D_i^2,$$

where b is the number of terminal nodes. At the root of the tree, the ''naive'' prediction is the overall mean $\bar{y}$. Thus, $D_r^2 = \sum_{i=1}^{b}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$ is the ''worst-case'' deviance. The relative error

$$E^2 = \frac{D^2}{D_r^2},$$

measures the improvement of the current model as compared with the simplest model. A reliable estimate of E^2 is obtained by cross validation (cf. Efron and Tibshirani 1993). Typically, the data are divided into v similar parts. The $v - 1$ parts are used as a training set for the tree-building process, and the remaining data serve as a test set for validating the tree. The overall fit is assessed by averaging the v estimates of error, obtaining the cross-validation relative error E_{cv}^2.

The analysis is based on the proprietary S-Plus implementation of CART (MathSoft 1998) in the tree function and its derivatives. Further description of this implementation can be found in Venables and Ripley (1994).

4. Fitting and validating the models

a. Constructing explanatory variables

Potential raw explanatory variables for the models are temperature, wind speed, wind direction, month, and hour of the day. Because the last three variables are cyclical, respective indicators were defined.

1) *Wind sector indicator w.* For the eastern wind sector, $w = -1$; $w = 1$ for the power plant sector; $w = 2$ for the refineries sector; and $w = 0$ for all other sectors. To determine the relevant wind sectors, we start by observing that the power plant is located in the 330° direction from the receptor. Koch and Dayan (1992) showed that, for the most frequent synoptic category (Persian trough with shallow pressure gradient, persisting in this mode about 2/3 of the days in the summer), a 5-yr mean anticyclonic veering of 4° (100 m)$^{-1}$ is measured in the atmospheric layer in which the plume has risen and been transported as released from a 300-m stack. The highest stacks in the studied region are 150 m, corresponding to a total veering of approximately 10°–15° in the ABL, where effective height of emitted plumes of such stacks are expected. Typical 30-min period standard deviations of the horizontal wind direction fluctuation (σ_θ) for summer noontime corresponding to a B–C Pasquill stability category is approximately 15°–20° (Gifford 1976). The standard deviation of concentration distribution in lateral direction (σ_y) for a 3-km downwind distance from the power plant under such stability conditions is 300–500 m (Slade 1968). These considerations lead us to conclude that the relevant wind direction indicating plumes originating from the power plant is within the sector 280°–340°. The refineries, located approximately in the 10° direction, have lower stacks than the power plant does and are closer to the receptor. In accord, the relevant sector assigned to this source is 350°–20°. The eastern sector is defined by the range 30°–180°.
2) *Time of day indicator h.* Between 0900 and 1800 LST, $h = 1$, and $h = 0$ otherwise. The daytime range was determined by the observed diurnal distribution of pollution episodes downwind of the power plant, which indicates a sharp decline before 0900 and after 1800, when convection induced by surface heating is suppressed.
3) *Season indicator m.* Between May and September, $m = 1$, and $m = 0$ otherwise, corresponding to the aforementioned power plant fuel policy.

b. Regression tree in original scale

Our objective is to classify RSN into categories characterizing distinct sources of pollution and to understand the principle sources of variation in RSN. It is therefore sufficient for our purposes to obtain a tree with a relatively small number of terminal nodes, reflecting the main sources of variation in RSN. We begin by restricting the minimal node size for splitting to 20 observations (i.e., growing continues if there are at least 20 observations in a node). The resulting initial tree had 30 terminal nodes with deviance $D^2 = 4953$. Figure 4 explores all subtrees of the initial tree by plotting the deviance against the number of terminal nodes. The figure indicates that the main gain in precision is obtained by the 9-nodes tree, with $D^2 = 5344$ and relative error $E^2 = 5344/6953 = 0.76$. The next tree for consideration has 16 nodes, and $D^2 = 5147$. The relatively small improvement in deviance does not seem to justify the added complexity (seven additional nodes). Figure 5 displays the resulting tree, and Table 1 characterizes the nodes. The salient features of the 9-nodes tree are as follows:

- *Level 0.* At the root, the overall mean RSN is 0.98.
- *Level 1.* The first split partitions the data by the wind sector indicator w, associated with the main stationary sources. On the right branch are observations for

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

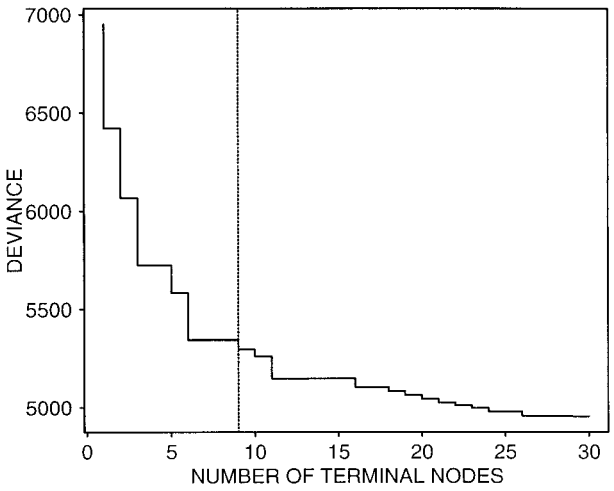

FIG. 4. Reduction in deviance as a function of the number of nodes for the 1996–97 data in original scale. The vertical line indicates the deviance for a 9-nodes subtree ($n = 5868$).

which $w > 1.5$, indicating the refinery direction ($n = 493$), and on the left branch are all other wind sectors. The mean RSN in the direction of the refineries is 2.00, as compared with 0.89 for all other observations.

- *Level 2.* At the second level, both nodes are partitioned by wind speed at thresholds of 3.75 and 2.75 m s^{-1} on the left- and right-hand sides, respectively. Mean RSN values are larger for faster wind episodes than for slower wind episodes (1.05 as compared with 0.71 on the left branch and 2.80 as compared with 1.05 on the right branch).
- *Level 3, nonrefineries sector.* A further split on the left branch reflects diurnal variation of RSN during faster wind episodes, with night and day means of 1.68 and 0.94, respectively. The relative error of the left branch is $E^2 = 0.91$.
- *Levels 3 and higher, refineries sector.* For the faster wind speed episodes, on the right-hand side, a further partitioning by temperature and wind speed is applied. The model identifies two small subgroups of observations ($n = 15$ and 10), with extremely high mean RSN (5.31 and 7.01 in nodes 7 and 9, respectively). It is clear that the RSN does not increase monotonically with temperature. The relative error of this branch is lower than that of the left branch ($E^2 = 0.73$). These findings are consistent with the descriptive analysis of section 2 and with the meteorological mechanism prevailing in the Ashdod airshed. The model adds insight into the pollution process by determining the cutoff points that define the underlying meteorological structure.

This tree illustrates the flexibility of the model in accommodating, for example, nonlinear relationships. It also shows that the same predictor may be used in several levels of the model. The asymmetry of the tree is evident, in that 6 of the 9 terminal nodes explain the structure of high values of RSN (which are only 8% of the observations). Because RSN values attributable to mobile sources are expected to be low, a further investigation of the variability of low values of RSN is carried out through a regression tree in logarithmic scale.

c. Regression tree in logarithmic scale

Similar criteria and considerations used for fitting the previous tree indicated a 9-nodes tree for the logarithm of RSN. The resulting tree is displayed in Fig. 6, and the nodes are defined and characterized in Table 1. The tree has an overall relative error of $E^2 = 0.73$ and is nearly symmetric. The first-level split partitions the data by wind speed, where again slower winds are associated with lower values of RSN. At the second level, spells of easterly winds are separated from those of noneasterly winds, with mean RSN of 0.47 and 0.74, respectively. On the right, the split reflects diurnal variation, with means of 0.97 and 1.74 in nighttime and in daytime, respectively. Note that the cited means are computed in the original scale.

d. Combined partition

The original- and log-scale trees define two partitions of the data into subsets (nodes). Denote these partitions by T_{org} and T_{log}, respectively, and the ith node of a tree T, by $T(i)$. The partition T_{log} focuses on the within-node homogeneity of low RSN values, and T_{org} focuses on homogeneity of high RSN values. Because both low and high values of RSN are of interest, we combine the two models by forming a new partition. The combined partition T_{com} is defined by the intersection of the subsets $\{T_{org}(i)\}$ and $\{T_{log}(i)\}$. Generally, this partition cannot be described by a binary tree. Nevertheless, its relative error $E^2(T_{com})$ can be computed by summing the squared error over all subsets. Because, for each pair of subsets $T_{com}(i)$ and $T_{org}(i)$, $T_{com}(i)$ is either included in $T_{org}(i)$ or the two subsets are disjoint, T_{com} is a refinement of T_{org} (Breiman et al. 1984, section 9.2). In a similar way, T_{com} is a refinement of T_{log}. Therefore, $E^2(T_{com}) \leq E^2(T_{org})$ and $E^2(T_{com}) \leq E^2(T_{log})$; that is, T_{com} performs at least as well as both its parent trees.

Analysis of the combined partition indicates that there are 22 nonempty subsets (out of 81 potential sets). The intersections of $T_{org}(2)$ and $T_{log}(5)$ and of $T_{org}(9)$ and $T_{log}(6)$ comprise two and four observations, respectively. These two subsets were omitted from the analysis. Table 2 displays the remaining 20 subsets, sorted by their mean RSN. As expected, the most prominent feature of this partition is that $T_{org}(1)$, consisting of low RSN values, intersects with essentially six nodes of T_{log}, and nodes $T_{org}(5)$–$T_{org}(9)$, consisting of the highest values of RSN, are collapsed into two nodes in T_{log}. In summary, the separate analyses based on T_{org} and T_{log} highlight the upper and lower tails of the distribution of RSN, respectively. In contrast, the combined parti-

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

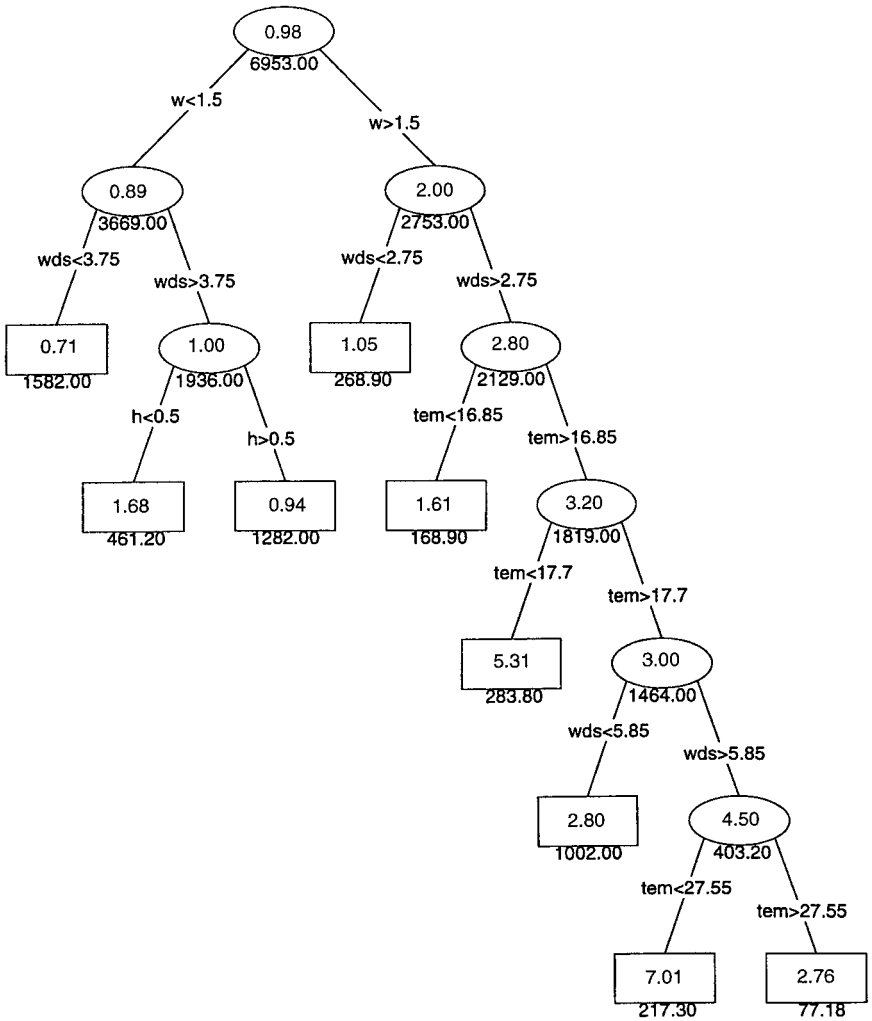

FIG. 5. A 9-nodes tree in original scale for 1996–97 (T_{org}). At each node, the mean value of RSN appears inside the ellipse (rectangle for terminal nodes) and the within-node deviance D_i^2 is beneath it. The split rule appears across the branches.

tion, based on the intersection of the two trees, incorporates the sources of variability in the full range of RSN values.

e. Linear regression model and cross validation

The last model we consider is the parametric multiple linear regression model. To comply with the linearity and variance homogeneity assumptions of the model, we use log(RSN) as the response variable. Because the results of CART indicate that the relationship between RSN and temperature is nonlinear, a second-degree polynomial in temperature was fitted to the data. Because the wind sector indicator receives four values, three dummy variables (w^*_{-1}, w^*_1, and w^*_2) were defined to represent this factor. In addition, interactions of $\{w_i^*\}$ with all other explanatory variables were tested for statistical significance. A stepwise regression algorithm was used to estimate the model for 1996–97. The resulting model, including all coefficients with significance level $P \leq 0.05$, in the order in which they entered the model, is given by

$$\begin{aligned}\hat{y} = &-1.32 + 0.17\text{wds} + 0.12(\text{wds})(w_2^*) - 1.25w^*_{-1} \\ &- 0.47h + 0.95(h)(w^*_{-1}) + 0.76(m)(w^*_{-1}) \\ &- 0.15(m)(w_1^*) + 0.43(h)(w_1^*) + 0.25(h)(w_2^*) \\ &- 0.05(\text{wds})(w_1^*) + 1.26w_1^* - 0.10(\text{tem})(w_1^*) \\ &+ 0.05\text{tem} - 0.001\text{tem}^2 + 0.002(\text{tem}^2)(w_1^*),\end{aligned}$$

where $y = \log(\text{RSN})$, wds is wind speed, and tem is temperature. The percent of explained variance is $R^2 = 0.30$ ($E^2 = 0.70$). Hence, the regression equation is consistent with CART's findings: wind speed plays a major role in determining RSN levels, with different slopes for the refineries and for the power plant wind sectors. RSN levels also vary with interactions of the

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

TABLE 1. Definition of nodes in T_{org} (Fig. 4) and T_{log} (Fig. 5) and distributional properties of RSN. Day: 0900–1800 LST; summer: May–Sep.

	Node definition					
Node	Wind sector	Wind speed (m s^{-1})	Other	n	Mean	Std dev
Original scale (T_{org})						
1	Nonrefineries	≤3.75		2586	0.71	0.78
2	Nonrefineries	>3.75	Night	407	1.68	1.07
3	Nonrefineries	>3.75	Day	2382	0.94	0.73
4	Refineries	≤2.75		226	1.05	1.09
5	Refineries	>2.75	tem ≤ 16.85°C	77	1.61	1.49
6	Refineries	>2.75	16.85° < tem ≤ 17.70°C	15	5.31	4.50
7	Refineries	2.75–5.85	17.70°C < tem	150	2.80	2.59
8	Refineries	>5.85	17.70° < tem ≤ 27.55°C	10	7.01	4.91
9	Refineries	>5.85	27.55°C < tem	15	2.76	2.35
Logarithmic scale (T_{log})						
1	East	≤2.75		364	0.47	0.77
2	Noneast	≤0.95	Nonsummer	107	0.33	0.32
3	Noneast	0.95–2.75	Nonsummer	870	0.67	0.68
4	Noneast	≤2.75	Summer	411	0.99	1.09
5	East	>2.75	Night	21	0.40	0.55
6	Noneast	>2.75	Night	698	1.79	1.64
7	Nonrefineries	2.75–4.55	Day	1713	0.83	0.79
8	Nonrefineries	>4.55	Day	1555	0.97	0.68
9	Refineries	>2.75	Day	129	2.75	2.85
All				5868	0.98	1.09

time of day and of the time of year with wind sector indicators. Temperature is the least influential factor.

To obtain a cross-validation estimate of the relative error, separate trees for 1996 and 1997 are constructed. The model constructed for 1996 is applied to 1997 data, and vice versa. The first and last rows in Table 3 display the ordinary estimates of relative error based on the training test. The relative error for the combined partition T_{com} was computed in both the original and logarithmic scales. It is seen that the models have similar cross-validation error estimates, ranging from 0.77 to 0.79 in the log scale and between 0.83 and 0.84 in the original scale. The cross-validation estimates are higher, on average, by 8%–10% than the respective ordinary estimates. Note that the linear and tree-based models have roughly the same predictive power, despite the fact that *categories* of temperature and wind speed are used in the tree-based model as opposed to the more detailed *continuous* values, used in the linear model. We conclude that, for the classification of meteorological conditions by their effect on RSN, the combined partition is more informative than the regression equation. We therefore use T_{com} in the subsequent analysis.

5. Results

a. Sources of variability in RSN

The 20 subsets of the combined partition reflect four underlying RSN levels, depicted in Table 2: very low (subsets 1–4), low (5–10), medium (11–15), and high (16–20). Figure 7 displays the distribution of observations in these four levels. The salient meteorological characteristics of observations included in each level are as follows:

- *Very low, interquartile RSN range [0.12, 0.48].* Observations in this level are characterized by easterly slow winds, not in the summer (highest frequency between November and January), and mainly in the cooler hours of the day (0600–1100 and 1500–0000 LST). Toward the end of autumn and most of the winter, when observed RSN in this level is the most frequent, 50%–60% of the winds blow from the easterly sector at a mean speed of 2.5 m s^{-1}. Being a marker for traffic-related pollution and characterized by a stable and persistent spatial pattern (Lebret et al. 2000), NO_x is probably trapped within the stable atmospheric surface layer formed by the rapid nocturnal radiative heat losses from the open agricultural area to the east of the receptor (see Fig. 1). Because morning and evening heavy traffic are the main sources of pollution east of the receptor, pollution events in this category are mainly attributable to mobile sources.
- *Low, interquartile range [0.43, 1.00].* Episodes in this category are typically in the western wind sector (43% in the power plant sector and the rest in the southwesterly sector), mainly during the day (0800–1800 LST), throughout the year. This category represents levels of RSN attributable to the power plant, because it includes 95% of the episodes downwind of this source. These are, apparently, indistinguishable from typical RSN values attributable to daytime emissions of stationary areal sources in the southwesterly sector.
- *Medium, interquartile range [0.74, 1.90].* Observations in this level occur on either episodes of south-

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

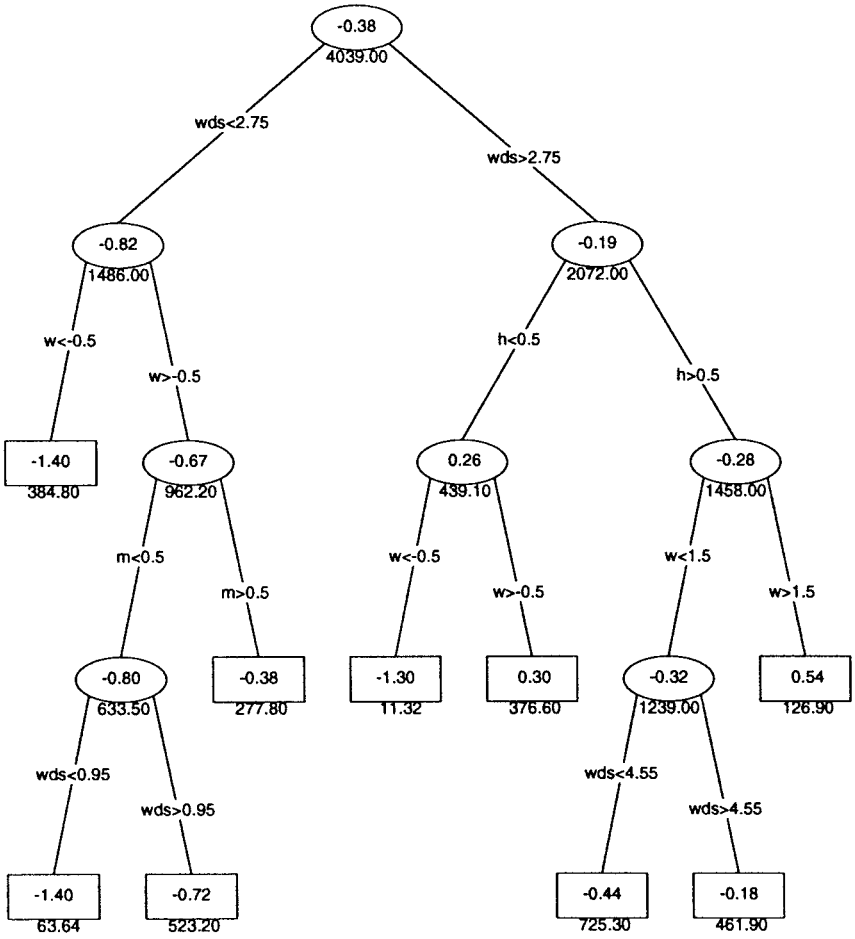

FIG. 6. A 9-nodes tree in logarithmic scale for 1996–97 (T_{log}).

TABLE 2. Distribution of RSN in the combined partition T_{com} defined by the intersection of terminal nodes of T_{org} and T_{log}.

Level	T_{com}-node	T_{org}-node	T_{log}-node	n	Mean	Std dev
Very low	1	1	5	17	0.28	0.19
	2	1	2	91	0.32	0.31
	3	4	2	16	0.40	0.42
	4	1	1	364	0.47	0.77
Low	5	1	3	742	0.62	0.61
	6	1	7	886	0.78	0.76
	7	3	7	827	0.88	0.83
	8	1	4	329	0.91	1.01
	9	3	8	1555	0.97	0.68
	10	4	3	128	0.98	0.94
Medium	11	1	6	157	1.19	0.95
	12	4	4	82	1.29	1.32
	13	5	9	29	1.58	1.47
	14	5	6	48	1.63	1.52
	15	2	6	403	1.69	1.06
High	16	7	9	75	2.69	2.71
	17	9	9	13	2.82	2.48
	18	7	6	75	2.91	2.49
	19	6	6, 9	15	5.31	4.50
	20	8	6, 9	10	7.01	4.91

westerly winds (peak at 260°–280°) or downwind of the refineries, mainly not during daytime, and on non-summer months (peak at March–April). Hence, medium levels of RSN are attributable to local industry during the cool hours of the day and to the refineries when the wind is slow. Surface wind distributions for March and April around 1900 LST (mode of the RSN distribution in this level) indicate that in March the most frequent wind directions are westerlies (11%) and northeasterlies (10%). In April, the prevailing winds are northerlies (21%) and northeasterlies (11%). This result explains the episodes observed of both natures: local industry affecting the receptor in the westerly wind regime and influence of the refineries during north-to-northeast flows.

- *High, interquartile range [1.07, 4.30].* This category comprises episodes of moderate to fast winds (>2.75 m s^{-1}), from the refineries sector, at temperatures exceeding 16.85°C, in late afternoon and early evening (1600–2100 LST), and during the transitional seasons. This may be explained by the high frequency (23%) of north-to-northeasterly winds characterizing the wind flow during the transitional seasons. Several fac-

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

TABLE 3. Ordinary and cross-validation estimates of relative error for T_{org}, T_{log}, T_{com}, and the linear regression model, by year.

Training set	Test set	Original scale		Logarithmic scale		
		T_{org}	T_{com}	T_{log}	T_{com}	Linear
1996	1996	0.76	0.73	0.70	0.68	0.69
	1997	0.82	0.81	0.80	0.78	0.79
1997	1996	0.85	0.84	0.77	0.76	0.76
	1997	0.73	0.73	0.72	0.69	0.70
E^2_{cv}		0.84	0.83	0.79	0.77	0.78

tors act in concert to reduce plume rise from the refineries sector, thus dispersing less efficiently their rich sulfur plumes, which leads to high RSN levels in the downwind sector: (a) slightly to moderately stable conditions, persisting during these seasons and time of day, which resist upward vertical motions; (b) moderate to strong winds, steering the refinery plumes to blow downwind, suppressing their initial rises; and (c) the relatively low temperatures of the emitted gases of these low stacks, which favor a rapid loss of buoyancy.

The next section assesses the contribution of different sources in the Ashdod area to the number of episodes of increased concentrations of SO_2 and NO_x during 1996–97.

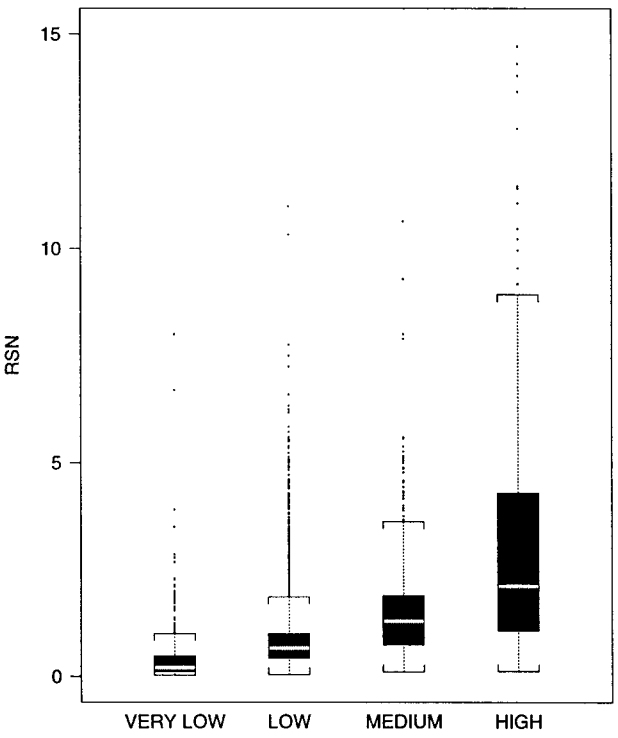

FIG. 7. Distribution of RSN by level, 1996–97. The box indicates the interquartile range, the center horizontal line is the median, and the vertical lines extend to 1.5 times the interquartile range. Dots above the staple represent outliers.

b. Analysis of episodes of increased concentrations

The findings of the previous section indicate that the SO_2/NO_x signature, in conjunction with wind direction, is a useful indicator of pollution sources. Analysis of RSN values for episodes with $SO_2 > 125$ μg m^{-3} or $NO_x > 235$ μg m^{-3} (one-quarter of the respective half-hourly air quality standards) shows that (a) for all wind sectors except the power plant sector, RSN values either do not exceed 0.37 or are higher than 0.53 and (b) daytime SO_2 episodes in the power plant sector, included in the low-level category, have a minimal RSN value of 0.49. However, for NO_x exceedances in this sector, RSN values are continuous, precluding a clear-cut distinction between traffic and power plant emissions. These findings yield the following source attribution rules: observations with RSN not exceeding 0.5 (approximately the third quartile of the very low level) were attributed to mobile sources, except for the power plant sector, for which the threshold was conservatively set to 0.1 (approximately the first quartile of the very low level). Observations with RSN larger than 0.5 were attributed to stationary sources except for the power plant, for which a threshold of 0.3 was used.

Although pollution events may result from a superposition of sources, episodes of increased SO_2 and NO_x concentrations might be frequently classified by a dominant source. Table 4 specifies the classification criteria and the estimated number of exceedances by source. Because wind direction is indicative of different stationary sources, intervals of 10°–15° separate the different sectors. However, pollution episodes attributable to mobile sources are identified mainly by their RSN fingerprint, and therefore the wind sector definitions for these sources are exhaustive. The results indicate that the respective numbers of SO_2 and NO_x exceedances in 1996–97 are 485 and 634. About 42% of the elevated SO_2 concentrations are attributed to the power plant, 33% to the refineries, 11% to local industry, and 14% to undefined sources. Most of the 66 SO_2 exceedances in the ''undefined'' group are in the 265°–279° sector, where it is unclear whether the power plant or local industry are accountable. The SO_2 1996 emissions inventory in the area indicates that, of the total emitted 36 343 Mg, 27 015 Mg (74.3%) were emitted by the power plant, 8010 Mg (22.0%) by the refineries, and the rest (3.7%) by local industry. Inventories for 1997

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

TABLE 4. Characterization of pollution sources and the number of exceedances beyond one-quarter of the respective air quality standards, 1996–97.

Source	Wind sector (°)	RSN threshold		SO_2 > 125 μg m^{-3}	NO_x > 235 μg m^{-3}
		Expected*	Observed**		
Elevated					
Power plant	280 ≤ wdd ≤ 340	>0.30	>0.30	204	29
Refineries	350 ≤ wdd, wdd ≤ 20	>0.50	>0.53	160	1
Areal-stationary					
East	30 ≤ wdd < 180	>0.50	>0.67	9	0
Southwest	180 ≤ wdd < 265	>0.50	>0.57	46	0
Areal–mobile					
East	20 < wdd < 180	≤0.50	≤0.28	0	458
Southwest	180 ≤ wdd < 270	≤0.50	≤0.34	0	56
Northwest	270 ≤ wdd < 340	≤0.10	≤0.10	0	24
North	340 ≤ wdd, wdd ≤ 20	≤0.50	≤0.37	0	37
Undefined				66	29
Total				485	634

* Predetermined cutoff point.
** Actual min/max values in the data.

had a similar distribution. As expected, the relative contribution of the power plant to the total emissions is higher than its relative contribution to ambient concentrations, and the inverse is true for the refineries. This result is explained by the higher effective height of the power plant stack as compared with the refinery stacks.

NO_x exceedances are mainly accounted for by traffic (91%), the power plant (4.5%), and undefined sources (4.5%). Most of the 29 NO_x exceedances in the undefined category have RSN between 0.1 and 0.3. No inventories for NO_x emissions are available for the study area. Comparison between 1996 and 1997 indicates that the number of elevated NO_x episodes slightly decreased (329 and 305 in 1996 and 1997, respectively). For SO_2, the number increased by approximately 150% (from 139 in 1996 to 346 in 1997).

Figure 8 shows the seasonal distribution of SO_2 exceedances. The lowest frequency is in the midsummer months of July and August. The highest impact of power plant emissions is in April and May and of the refineries is in April and October–November. Areal stationary sources also have the largest contribution in April. Traffic emissions, manifested in high NO_x episodes, have the largest effect between November and January (69% of the exceedances) and almost no effect during the summer.

6. Summary and conclusions

The paper proposes a method for studying the source–receptor relationship when regular emission inventories are missing or incomplete. The usefulness of the SO_2/NO_x signature in distinguishing between pollution sources was demonstrated for the southern coastal plain of Israel, where diverse pollution sources affect air quality. The sources of variability of RSN were analyzed by tree-based regression models and by a linear model. The main advantage of the tree-based procedures proved to be the insight gained by the unique subset structure that they defined. The combination of two trees, in original and logarithmic scales, enabled a simultaneous investigation of the low and high RSN values. The resulting combined partition highlighted the underlying meteorological conditions, which yield different levels of RSN. The agreement between the combined partition and the inferential findings of the linear regression model strengthened the analysis. Furthermore, although tree-based predictions are constant for all observations in the same subset, the relative prediction error of CART was similar to that of the linear model.

The explanatory power of the variables at hand was low, as suggested by the fact that the relative error of a ''near-perfect'' tree (912 nodes) is $E^2 = 0.51$. Clearly, more accurate meteorological data based on synoptic conditions, depth of the mixing layer, ventilation rates, and other ABL parameters are expected to improve the model performance. This area will be dealt with in future work.

The results show that RSN identifies mobile sources for all wind sectors except for the power plant sector. For stationary sources, the RSN signature should be supplemented by wind direction.

It is interesting that, in earlier years, RSN values in the power plant sector were indistinguishable from those in the refineries sector. The decrease in the power plant SO_2/NO_x signature is a direct result of the improvement in fuel quality and of the control policy.

Classification of episodes of increased SO_2 and NO_x concentrations by their probable source indicate that power plant SO_2 emissions are controlled successfully during the summer but not during winter and spring and that refineries' SO_2 emissions have a prominent effect on air quality during the transitional seasons. NO_x emissions are mostly attributable to mobile sources. It is

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

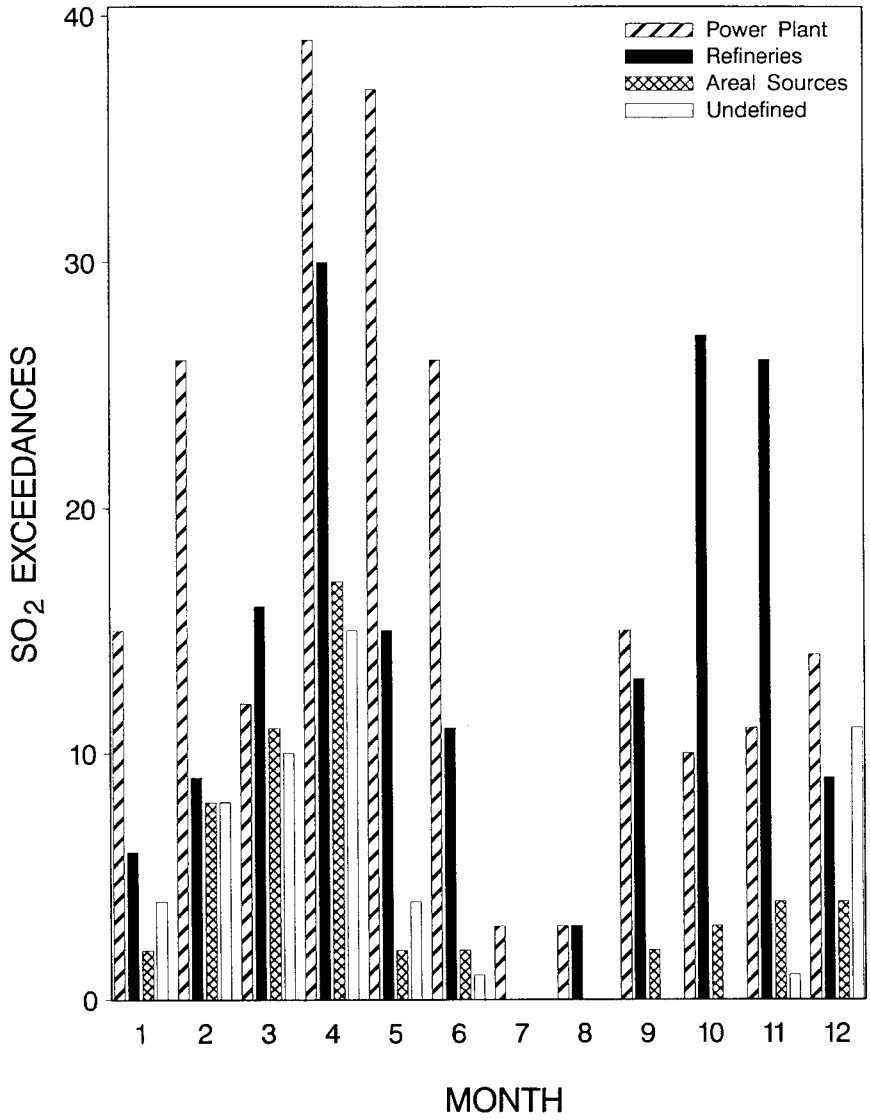

FIG. 8. Number of exceedances of SO_2 beyond one-quarter of the air quality standard (125 μg m^{-3}) by pollution source and month, 1996–97 (n = 485).

regrettable that the number of NO_x exceedances beyond one-quarter of the air quality standard nearly tripled between 1995 and 1996, indicating that traffic emissions have become a major air quality problem in the study area.

Acknowledgments. The authors thank Doron Lahav of the Ashdod–Hevel Yavne Association of Towns for the Protection of Environment Quality for providing the data and Michal Kidron from the Cartographic Laboratory of The Hebrew University for her kind assistance in the preparation of the figures. The authors also thank the referees for their valuable comments, which led to significant improvement in the paper. This research was partially funded by the Ministry of the Environment, Israel.

REFERENCES

Balmor, Y., A. Gutman, and U. Dayan, 1988: Wind and temperature structure of the atmospheric boundary layer in a coastal versus inland site in Israel. *Proc. EURASAP Conf: Meteorology and Atmospheric Dispersion in a Coastal Area,* Riso, Denmark, EURASAP, 15–24.

Benkovitz, C. M., and Coauthors, 1996: Global gridded inventories of anthropogenic emissions of sulfur and nitrogen. *J. Geophys. Res. Atmos.,* **101,** 29 239–29 253.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, 1984: *Classification and Regression Trees.* Wadsworth and Brooks, 358 pp.

Burrows, W. R., 1997: CART regression models for predicting UV radiation at the ground in the presence of cloud and other environmental factors. *J. Appl. Meteor.,* **36,** 531–544.

——, M. Benjamin, S. Beauchamp, E. R. Lord, D. McCollor, and B. Thomson, 1995: CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Van-

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC

couver, Montreal, and Atlantic regions of Canada. *J. Appl. Meteor.,* **34,** 1848–1862.

Carter, M. M., and J. B. Elsner, 1997: Statistical method for forecasting rainfall over Puerto Rico. *Wea. Forecasting,* **12,** 515–525.

Dayan, U., and J. Koch, 1989: Assessment of the critical conditions for dispersion and transport of plumes from tall stacks in the Haifa area. *Proc. Fourth Int. Conf. of the Israel Society for Ecology and Environmental Quality Sciences*, Jerusalem, Israel, ISEEQS, 27–36.

——, R. Shenhav, and M. Graber, 1988: The spatial and temporal behavior of the mixed layer in Israel. *J. Appl. Meteor.,* **27,** 1382–1394.

Duncan, B. N., A. W. Stelson, and C. S. Kiang, 1995: Estimated contribution of power plants to ambient nitrogen-oxides measured in Atlanta, Georgia in August 1992. *Atmos. Environ.,* **29,** 3043–3054.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap.* Chapman and Hall, 436 pp.

Gifford, F. A., 1976: Turbulence diffusion typing schemes—A review. *Nucl. Saf.,* **17,** 68–86.

Hopke, P. K., 1985: *Receptor Modeling in Environmental Chemistry.* John Wiley and Sons, 319 pp.

Koch, J., and U. Dayan, 1992: A synoptic analysis of the meteorological conditions affecting dispersion of pollutants emitted from tall stacks in the coastal plain of Israel. *Atmos. Environ.,* **26A,** 2537–2543.

Lebret, E., and Coauthors, 2000: Small area variations in ambient NO_2 concentrations in four European areas. *Atmos. Environ.,* **34,** 177–185.

MathSoft, 1998: *S-PLUS 5 for UNIX Guide to Statistics.* Data Analysis Division, MathSoft, 1014 pp.

Ryan, W. F., 1995: Forecasting severe ozone episodes in the Baltimore metropolitan-area. *Atmos. Environ.,* **29,** 2387–2398.

Slade, D. H., 1968: *Meteorology and Atomic Energy.* Office of Information Services, U.S. Atomic Energy Commission, 444 pp.

Swietlicki, E., S. Puri, H.-C. Hansson, and H. Edner, 1996: Urban air pollution source apportionment using a combination of aerosol and gas monitoring techniques. *Atmos. Environ.,* **30,** 2795–2809.

Venables, W. N., and B. D. Ripley, 1994: *Modern Applied Statistics with S-Plus.* Springer, 462 pp.

Walmsley, J. L., W. R. Burrows, and S. Schemenauer, 1999: The use of routine weather observations to calculate liquid water content in summertime high-elevation fog. *J. Appl. Meteor.,* **38,** 369–384.

Watson, J. G., N. F. Robinson, J. C. Chow, R. C. Henry, B. M. Kim, T. J. Pace, E. L. Meyer, and Q. Nguyen, 1990: The USEPA/DRI Chemical Mass Balance Receptor Model, CMB7.0. *Environ. Software,* **5,** 38–49.

Unauthenticated | Downloaded 12/06/21 06:51 PM UTC