Environ Monit Assess
DOI 10.1007/s10661-013-3101-y

# Statistical analysis of aerosol species, trace gasses, and meteorology in Chicago

**Katrina Binaku · Timothy O'Brien · Martina Schmeling · Tinamarie Fosco**

Received: 13 June 2012 / Accepted: 16 January 2013


**Abstract** Both canonical correlation analysis (CCA) and principal component analysis (PCA) were applied to atmospheric aerosol and trace gas concentrations and meteorological data collected in Chicago during the summer months of 2002, 2003, and 2004. Concentrations of ammonium, calcium, nitrate, sulfate, and oxalate particulate matter, as well as, meteorological parameters temperature, wind speed, wind direction, and humidity were subjected to CCA and PCA. Ozone and nitrogen oxide mixing ratios were also included in the data set. The purpose of statistical analysis was to determine the extent of existing linear relationship(s), or lack thereof, between meteorological parameters and pollutant concentrations in addition to reducing dimensionality of the original data to determine sources of pollutants. In CCA, the first three canonical variate pairs derived were statistically significant at the 0.05 level. Canonical correlation between the first canonical variate pair was 0.821, while correlations of the second and third canonical variate pairs were 0.562 and 0.461, respectively. The first canonical variate pair indicated that increasing temperatures resulted in high ozone mixing ratios, while the second canonical variate pair showed wind speed and humidity's influence on local ammonium concentrations. No new information was uncovered in the third variate pair. Canonical loadings were also interpreted for information regarding relationships between data sets. Four principal components (PCs), expressing 77.0 % of original data variance, were derived in PCA. Interpretation of PCs suggested significant production and/or transport of secondary aerosols in the region (PC1). Furthermore, photochemical production of ozone and wind speed's influence on pollutants were expressed (PC2) along with overall measure of local meteorology (PC3). In summary, CCA and PCA results combined were successful in uncovering linear relationships between meteorology and air pollutants in Chicago and aided in determining possible pollutant sources.

**Keywords** Atmospheric aerosols · Canonical correlation analysis · Chicago air pollution · Multivariate statistics · Principal component analysis · Trace gasses

K. Binaku · M. Schmeling (✉) · T. Fosco
Department of Chemistry, Loyola University Chicago,
1032 West Sheridan Road,
Chicago, IL 60660, USA
e-mail: mschmel@luc.edu

T. O'Brien
Department of Mathematics and Statistics, Loyola University Chicago,
1032 West Sheridan Road,
Chicago, IL 60660, USA

## Introduction

Many air pollution studies involve collection and analysis of atmospheric aerosols and concurrent meteorology measurements in regional areas to identify a

Published online: 01 February 2013

region's pollutant signature (Fosco and Schmeling 2006, 2007; Shen et al. 2009). The composition of aerosols is region-specific and encompasses inorganic and organic species of natural and anthropogenic origin, present due to primary emission or secondary formation in the atmosphere. Aerosol sampling campaigns result in substantial databases containing regional air pollutant concentrations in particulate and/or trace gas form along with meteorological data corresponding to sampling duration. Maximizing the information these data sets reveal enhances the benefit of completing timely aerosol collection and pollutant research campaigns.

One major reason for detailed studies of aerosol particles and trace gasses is the link between respiratory ailments and pollutant ($PM_{10}$, $PM_{2.5}$, trace gasses) levels. Schlesinger (2007) reviewed a plethora of studies linking short- and long-term respiratory ailments to the inorganic particulate portion of $PM_{2.5}$. The author pointed out that constituents in aerosols vary regionally, and, as a result, health studies differ in conclusions of whether or not certain pollutants have detrimental effects (Schlesinger 2007). Studies involving controlled ozone exposure found that ozone caused inflammation of lung tissue and temporary lung function reduction in humans (Bascom et al. 1996). As air pollutants can exist in the atmosphere from hours to several days, transport of pollutants from emission sources likely impacts regions far from the original source. Therefore, determining relationships between meteorology and air pollutants is important to predict pollution episodes and to identify regional emission sources.

Multivariate statistical techniques are applied to atmospheric data to reduce data dimensions as well as aid in determining natural and anthropogenic sources of air pollutants. Principal component analysis (PCA) and factor analysis (FA) have been applied to air pollutant and meteorological data from all over the world to identify sources of pollutants (inorganic ions, heavy metals, or trace gasses) monitored in a particular study (Almeida et al. 2006; Braga et al. 2005; Harrison et al. 1996; Hsieh et al. 2008; Ravindra et al. 2008; Statheropoulos et al. 1998). Canonical correlation analysis (CCA) has been used to evaluate atmospheric data considerably less than PCA/FA and for a different purpose: to determine linear relationships between air pollution variables and meteorological parameters (Braga et al. 2005; Statheropoulos et al. 1998). Uncovering these relationships can allow researchers to gauge pollution transport, pollutant sources, and better predict local pollution episodes.

Chicago is a large Midwestern city with roughly 2.7 million residents and including the surrounding metropolitan area is comprised of about 9.5 million residents in total (U.S. Census Bureau 2010). Several interstate highways link the Chicago metropolitan area to other cities in the region (Simcik et al. 1999) and to two major airports, O'Hare International Airport and Midway International Airport. Major point sources in Chicago include two coal-fired power plants along with metal processing, paint, and solvent factories in the south and Southeast areas of the city. Coke ovens, steel manufacturing facilities, and oil refineries are several point source polluters in Northwest Indiana (Simcik et al. 1999). Emissions from the mentioned point and mobile sources contribute to the atmospheric pollution signature of the Chicago region substantially (USEPA 2010). The composition of PM in the Midwest is primarily sulfate and organic carbon species, followed by secondary nitrate (USEPA 2010). Lee et al. (1993) evaluated Chicago pollution data and determined that a majority of nitrogen compounds result from mobile source emissions, while local sulfur compounds are a result of a combination of refinery, coal, and steel manufacturing. Regional transport of pollutants also contributes to the local atmospheric signature. Scheff et al. (1984) found that winds originating from the south transported air masses comprised of higher particle concentrations of sulfate and nitrate to Chicago than winds originating from the north. Studies in another Midwestern city, a designated USEPA supersite, St. Louis, MO, also suggest that regional transport of sulfate and other secondary species is a significant contribution to urban air pollution in the Midwest (Lee and Hopke 2006).

In this study, both CCA and PCA were applied to atmospheric data collected in Chicago, IL, during summers 2002, 2003, and 2004. Data include various air pollutant concentrations and several meteorological parameters. The purpose of applying multivariate statistical techniques to the data set was to determine the extent of linear relationships between aerosol concentrations, trace gas mixing ratios, and meteorological parameters in addition to reducing data dimensions and determining sources of local air pollutants.

*Canonical correlation analysis* is a multivariate statistical method that determines the extent of existing

linear relationships, or lack thereof, between two sets of data containing multiple variables in each (Hair et al. 1992). Generally, one data set corresponds to variables defined as independent while the other data set contains variables classified as dependent (Hair et al. 1992). For each data set, linear combinations are derived; canonical weights within linear combinations are generated in such a way that maximum correlation is achieved between the linear combinations of the first data set and the linear combinations of the second data set (Hair et al. 1992; Johnson and Wichern 1998). The first canonical function derived reflects the maximum linear correlation possible between the two original data sets. Each successive canonical function derived maximizes residual inter-correlations between data sets not explained by previous canonical functions. Therefore, each canonical variate pair is orthogonal and uncorrelated to one another (Hair et al. 1992).

Shown below are the general equations (adapted from Manly 2005) for linear combinations of data set no. 1 ($U$) and data set no. 2 ($V$); combined, they represent a canonical function ($U$, $V$).

$$U = \mathrm{a}_{11}X_1 + \mathrm{a}_{12}X_2 + \mathrm{a}_{13}X_3 + \mathrm{a}_{14}X_4 \ldots\ldots a_{1n}X_n \quad (1)$$

$$V = \mathrm{b}_{11}Y_1 + \mathrm{b}_{12}Y_2 + \mathrm{b}_{13}Y_3 + \mathrm{b}_{14}Y_4 \ldots\ldots b_{1q}Y_q \quad (2)$$

In Eq. 1, symbols $X_1$, $X_2$, $X_3$, $X_4 \ldots X_n$ represent each original variable within data set no. 1; $a_1$ denotes the canonical weights in the linear combination ($U$) and "$n$" distinguishes that each weight value is different for each variable. In Eq. 2, $Y_1$, $Y_2$, $Y_3$, $Y_4 \ldots Y_q$ represent each original variable within data set no. 2; $b_1$ stands for each canonical weight in linear combination ($V$), and "$q$" distinguishes that each weight value is different for each variable. All successively derived canonical functions follow the equations displayed above.

Within each canonical function are canonical weights corresponding to the amount of influence each original variable has in the linear combination (Hair et al. 1992; Manly 2005). The larger the weight, the more influence a particular variable associated with the weight value has in the linear combination. In addition to canonical functions, CCA also derives several other important pieces of information: (1) canonical correlations corresponding to the linear correlation between each derived canonical function, (2) statistical significance of each canonical function, and (3) simple correlations between original variables and respective derived canonical variates (Hair et al. 1992). Terminology of CCA results varies by reference; therefore, interchangeable terms are listed for clarity in Table 1. Additionally, linear combinations are used to calculate canonical scores, which project original variables' observations in the canonical function space. These plots can be examined to identify extreme cases, outliers, and trends (Hair et al. 1992).

Traditionally, canonical weights are the CCA result most interpreted through their magnitude and size (Hair et al. 1992). However, interpretation of canonical loadings, simple correlations between original variables, and respective canonical variates, is also done. Canonical loadings reveal additional information when original variables display colinearity or multicolinearity between themselves (Hair et al. 1992; Manly 2005). Colinearity is defined as the correlation between two variables; multicolinearity refers to multiple variables having correlation with one another (Manly 2005). If original variables exhibit colinearity prior to CCA, canonical weights can be misleading (Hair et al. 1992). It is important to note the absence or presence of colinearity of original variables to determine whether canonical weights, loadings, or both, should be interpreted. See Johnson and Wichern (1998) for in-depth derivation of the mentioned CCA component results.

*Principal component analysis* functions as a statistical method for dimension reduction of a multivariate set of data. Through PCA application, original variables are transformed into new variables; this transformation is performed to retain the variance of the original data but express this variability in a fewer number of new variables, thus eliminating redundant information (Wilks 2011). The new variables are expressed as linear combinations of original variables (Wilks 2011). These linear combinations are referred to as principal components (PCs) or eigenvectors. The

**Table 1** Summary table of CCA terminology of synonymous association

| Main term | Synonymous term(s) |
|---|---|
| Linear combination | Canonical variate |
| Canonical function | Canonical variate pair |
| | Canonical variable |
| Canonical weight | Canonical coefficient |
| Canonical loading | Canonical structure loading |

coefficients or loading values within a linear combination can be used to interpret relationships across a set of original variables and for classification purposes (Wilks 2011).

By rotation of the axes the original data occupy, loading values within principal components are generated to maximize the amount of variance explained by the new variable. The first PC captures the largest variance of the original data and is also associated with the largest eigenvalue (Wilks 2011). The second PC derived lies orthogonal to the first and captures variance not explained by the first PC. Each successive PC maximizes the data variance not expressed by the PCs preceding it. All PCs derived are orthogonal and uncorrelated with one another (Wilks 2011). Furthermore, the eigenvalue associated with each PC becomes smaller with each successive new linear combination. Eigenvalues aid in determining how many PCs are retained for interpretation. When following Kaiser's Rule, all principal components with an associated eigenvalue of <1.0 are not interpreted because the information gained is insignificant (Wilks 2011). Another tool to determine the number of PCs to retain is a Scree plot, a graph of eigenvalue versus corresponding PC number (Johnson and Wichern 1998; Wilks 2011). Consideration of a large change in slope between points in a Scree plot is used to determine which PCs to retain (Johnson and Wichern 1998; Wilks 2011). PC eigenvalues in the area of large slope are retained, whereas values located where the plot's curve levels off are not considered.

PCA is performed on either the covariance or correlation matrices of the original data. A correlation matrix is chosen when variables within the original data set were measured on varying scales; variables are standardized before PCA is applied (Wilks 2011). Similarly to CCA, PCA score plots projecting PC scores in new variable space are generated for interpretation. Additional information on PCA can be found in Johnson and Wichern (1998) and also Wilks (2011).

## Experimental details

### Sample collection and analysis

On weekdays during the summer months of 2002, 2003, and 2004, aerosol samples were collected and meteorology/trace gasses were monitored in Chicago, IL. As a coastal city bordering Lake Michigan, Chicago is susceptible to lake breezes, especially during summer months when the temperature difference between the lake and bordering land is large. Aerosol collections were carried out in the summer to study the difference in atmospheric signature between lake breeze events and non-lake breeze days. The hypothesis is that lake breeze events may impose short-lived pollution episodes over Chicagoland (Fosco and Schmeling 2006, 2007).

Loyola University Chicago served as the air sampling site for the mentioned pollution studies. The university is located in a residential area and is 13 km north of the city's center. Lake Michigan borders the university campus to the east (Fosco and Schmeling 2007). Instrumentation including aerosol collectors, trace gas monitors, and a weather station was operated atop Mertz Hall, a residential building 60 m in height and located 200 m to the west of Lake Michigan. A map displaying the aerosol collection sampling station in relation to downtown Chicago and surrounding suburbs is presented in Fig. 1.

During each summer study, sampling of aerosols took place on weekdays in two segments per day: a segment of collection from 0700 to 1000 hours and a second collection segment from 1100 to 1300 hours local time. These segments were designed to capture aerosols prior to and during a lake breeze event as well as measure formation of secondary aerosol species from morning to afternoon (Fosco and Schmeling 2007). Aerosol samples were collected on pre-cleaned 47-mm quartz fiber filters. The pre-cleaning procedure for the quartz filters is described in Fosco and Schmeling (2006, 2007) along with more details regarding sampling. Meteorological parameters monitored included: temperature (in degrees Celsius), wind speed (per millisecond), wind direction (in degrees), and humidity (in percent). The data were recorded every 15 min using a Vantage Pro$^{TM}$ Weather Station (Davis Instruments, Hayward, CA) mounted adjacent to the aerosol collection equipment. Concurrently to aerosol collections and recording meteorology, mixing ratios of ozone ($O_3$) and nitrogen oxides ($NO_x$) were monitored continuously (24 h) with 1-min resolution using a Thermo 49C $O_3$ Analyzer and a Thermo 42C NO-$NO_2$-$NO_x$ Analyzer (Thermo Environmental Instruments, Inc., Franklin, MA).

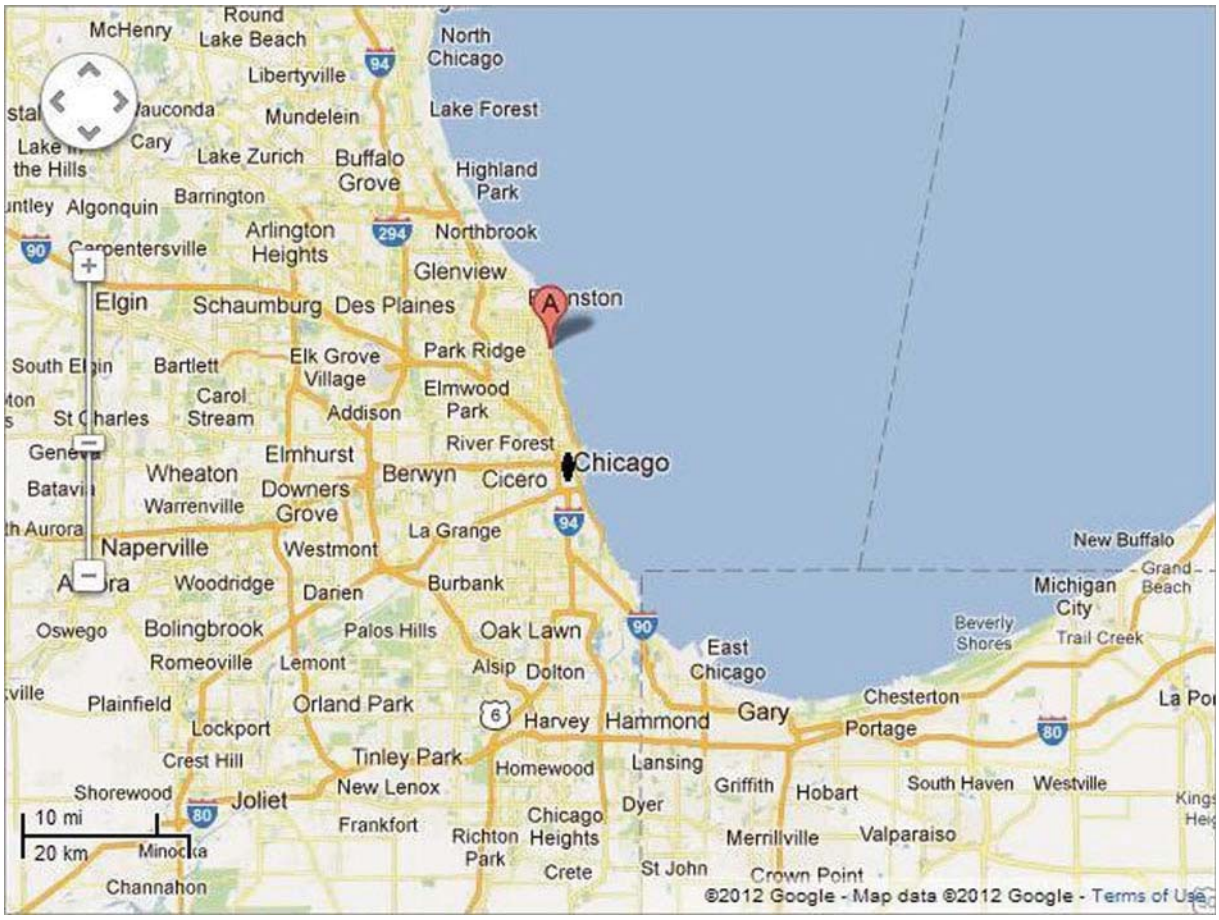



**Fig. 1** Map of Chicagoland including the aerosol sampling site (*A*) in relation to the downtown area of Chicago denoted by a *black circle* (Google 2012)

Following water extraction of aerosol material collected on quartz fiber filters, water soluble cationic and anionic aerosol species were quantified using a Metrohm 761 Compact Ion Chromatograph with chemical suppression (Metrohm USA, Inc., Riverview, FL). Cationic species present in aerosol samples and included in statistical analysis were ammonium ($NH_4^+$) and calcium ($Ca^{2+}$). Anionic species present in collected aerosol samples and included in statistical analysis were nitrate ($NO_3^-$), sulfate ($SO_4^{2-}$), and oxalate ($C_2O_4^{2-}$). All other ions were below detection limit or poorly resolved.

## Data standardizing and organization

Since all of the variables' values in the original data had varying levels of resolution due to differing instrument sampling frequency, meteorological data and trace gas mixing ratios were averaged to match the aerosol sampling segments. The final dataset consisted of 110 observations and 11 variables.

All variables were standardized by subtracting each variable's observed value by its respective average and then dividing by its standard deviation. Concentrations below detection limits were not considered in contrast to other studies published using multivariate statistical techniques which substituted detection limits or averages for missing data values prior to statistical application (Statheropoulos et al. 1998; Yu and Chang 2006; Zhou et al. 2007).

The statistical program used to perform CCA was SAS® 9.2 Software (SAS Institute Inc 2010). The results of CCA presented in SAS included testing for significance of canonical functions as well as derived canonical weights, canonical loadings, and canonical scores for each canonical function. PCA was applied

to data using Minitab® 16 Statistical Software (Minitab Inc 2010) and results included eigenvalues, eigenvectors, and principal component scores. Score plots for both CCA and PCA results were also generated using Minitab® 16 Statistical Software.

The original data, 110 observations of 11 variables (four meteorological parameters and seven air pollutants), were arranged in one data matrix (for PCA) and split into two data matrices (for CCA): the first containing concentrations (in microgram per cubic meter) of measured air pollutants (ammonium, A; calcium, C; nitrate, N; sulfate, S; and oxalate, Ox—along with mixing ratios of ozone, $O_3$; and nitrogen oxides, $NO_x$) and the second containing observed meteorological parameters during air sampling (wind speed, Sp; temperature, T; wind direction, D; and humidity, H). Throughout the results section, the original variables are denoted by the symbol associated with them as presented above. The meteorological parameters were considered predictor/independent variables, and air pollutants were considered response/dependent variables.

## Results

### Canonical correlation analysis

#### *Canonical variate pairs*

Four canonical functions were derived in CCA. The first three functions were statistically significant at the 0.05 level; therefore, the fourth canonical function was not reported since its interpretation would not add a significant value. Linear combinations derived corresponding to meteorological parameters are denoted Met1, Met2, and Met3; whereas linear combinations corresponding to air pollutants are denoted A.Poll1, A.Poll2, and A.Poll3. The canonical correlation of each canonical function (Met no., A.Poll no.) as well as the canonical weights corresponding to respective air pollution and meteorology linear combinations is displayed in Table 2.

The first canonical variate pair had a canonical correlation of 0.821, indicating that A.Poll1 and Met1 are highly correlated. Furthermore, 72.0 % of the variance in A.Poll1 is explained by Met1. A.Poll1, the first air pollutant linear combination, is mostly influenced by ozone (0.9004) and to a much lesser extent by oxalate (0.2259) and nitrogen oxides (0.1479). The canonical weights of the other variables in A.Poll1 are near zero. Temperature (0.9012) is the dominant independent variable in the first linear combination, Met1, corresponding to weather parameters. Therefore, according to the first canonical variate pair, the mixing ratio of ozone is positively correlated to temperature, whereas nitrogen oxides mixing ratio has low correlation. Thus, with rising temperature, ozone mixing ratios are increasing at a much larger rate as opposed to nitrogen oxide mixing ratios. It is well documented that ozone formation increases with temperature, mostly due to the photolytic destruction of $NO_x$. Hence, an anti-correlation between the two species is reflected in the first canonical variate pair by the contrast in their respective canonical weight values.

**Table 2** Standardized canonical weights for air pollutants (A.Poll) and meteorological parameters (Met) canonical variables along with canonical correlations for each canonical variate pair (Met no., A.Poll no.)

| Canonical function | Met1, A.Poll1 | Met2, A.Poll2 | Met3, A.Poll3 |
|---|---|---|---|
| Canonical correlation | 0.821 | 0.562 | 0.461 |
| Wind speed | 0.2025 | 0.5520 | −0.7792 |
| Temperature | 0.9012 | 0.3689 | 0.5816 |
| Wind direction | −0.2970 | −0.0576 | 0.6184 |
| Humidity | −0.1086 | 0.9004 | 0.4857 |
| Ammonium | −0.1116 | 0.8191 | 0.7137 |
| Calcium | 0.0999 | 0.1194 | −0.2732 |
| Nitrate | −0.0045 | −0.1267 | 0.5261 |
| Sulfate | 0.0349 | −0.1193 | −0.5433 |
| Oxalate | 0.2259 | 0.2009 | 0.4976 |
| $O_3$ | 0.9004 | −0.8350 | −0.5608 |
| $NO_x$ | 0.1479 | −1.2610 | 0.0999 |

Also interpreted through the canonical weights of the first function is the positive relationship between oxalate, the anion of oxalic acid, and temperature. The concentration of oxalate, however, is presumed to be of low correlation due to the small value of its weight. This is the extent of interpretable information contained in the first canonical variate pair. The top graph in Fig. 2 is a plot of the canonical scores generated by A.Poll1 and Met1. Overall, no yearly trends are observed, nor are there obvious outliers; the positive linearity of canonical scores can be seen, numerically stated earlier by the canonical correlation value (0.821).

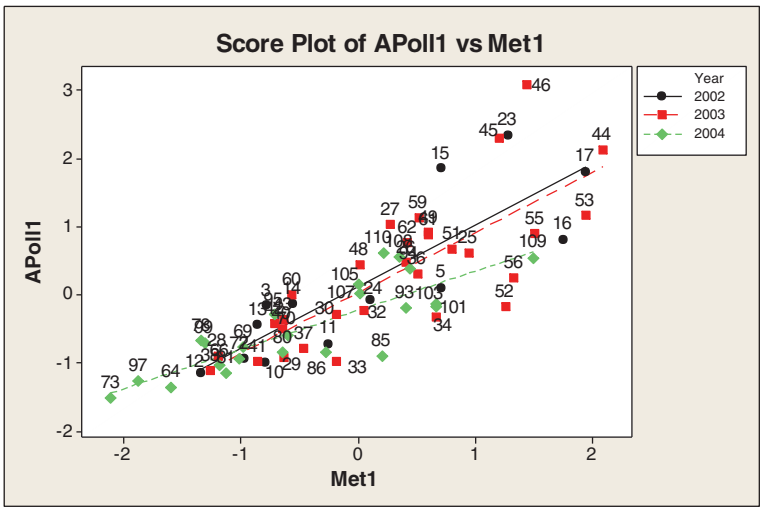
The canonical correlation of the second canonical variate pair (Met2, A.Poll2) was 0.562, indicating that there is moderate linear correlation between canonical variates. The amount of variance in A.Poll2 that is explained by Met2 is 16 %. Ammonium (0.8191), nitrogen oxides (−1.2610), and ozone (−0.8350) residuals not explained by the first canonical variate have the largest canonical weights and influence on A.Poll2. The weight of ammonium is positive, while the weights of both $O_3$ and $NO_x$ are negative, implying the presence of an inverse relationship. The weights of wind speed and humidity, 0.5520 and 0.9004, respectively, indicate that these meteorological variables have the most influence on Met2. This canonical variate pair suggests that moderate/high winds and very humid conditions are predictors of large ammonium concentrations. The second function can therefore distinguish humid/non-humid days as well as differences in observed wind speed and its effect on $NH_4^+$ concentrations and $NO_x$ mixing ratios. The large negative canonical weight value associated with $NO_x$ indicates that on days of high humidity and moderate winds, mixing ratios of $NO_x$ are low. This could be a result of increased mixing of air masses, which also explains the large negative weight associated with ozone. While the small canonical weight associated with wind direction implies little to no influence of wind direction on Met2, it is important to note that to the west/northwest of the sampling site areas with considerable agricultural signature are present. This may influence measureable ammonium in the area, increasing concentrations when the origin of the wind direction is from the west/northwest.

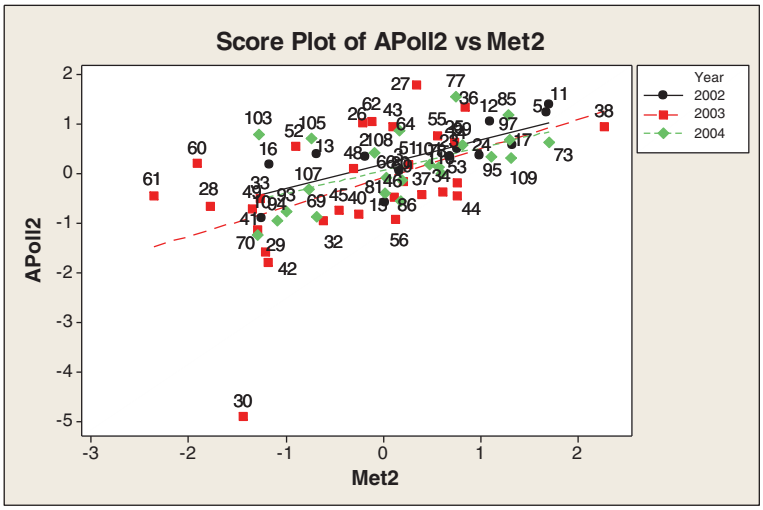
There is an outlier present in the A.Poll2 vs. Met2 score plot, the second graph in Fig. 2. An observation (no. 30) during summer 2003 is highly negative on A.Poll2 in relation to the other canonical scores. This indicated that one or more of the original variables with corresponding large negative canonical weight had substantial influence on the observation's calculated score. Upon further inspection of observed values in the original data set, the outlier occurred due to unusually high mixing ratios of $NO_x$ (negatively weighted in A.Poll2) measured during and after the normal traffic rush hour period on July 1, 2003. The large negative canonical weight of NOx on A.Poll2 directly contributed to the large negative score for the observation. Furthermore, winds originated from the southwest (202.5°) and turned south–southeast (157.5°). Both are directions in which heavily industrialized areas are located. While wind direction is not weighted significantly in Met2, it may have played a role in the large $NO_x$ observed, resulting in the outlier score.

The third statistically significant canonical function derived in CCA had a canonical correlation of 0.461. This canonical function represents residual intercorrelations which were not expressed by the first two canonical functions. Roughly 9.0 % of the variance in A.Poll3 is explained by Met3. Therefore, even though it is statistically significant, it does not reveal much new information. Wind direction is the only information provided in Met3 that was not significantly weighted in the previous two canonical functions. All residuals of the secondary pollutants have significant weights of varying sign, but interpretation is not clear since these are residual values of low canonical correlation. The bottom score plot in Fig. 2 for the third canonical function supports the lack of potential in interpreting the canonical variate pair.

### *Canonical loadings*

Canonical loadings were interpreted in addition to canonical weights to uncover additional information regarding relationships between air pollutants and meteorology. Canonical loadings represent the simple correlation between original variables and their respective canonical variate whereas canonical weights are values derived for each variable that maximize linear correlation between data sets (Hair et al. 1992; Johnson and Wichern 1998). A correlation matrix (Table 3) of the original data unveiled colinearity of several air pollutants and low to moderate correlations between meteorological parameters. Ammonium and calcium are both positively correlated with nitrate, 0.699 and 0.664, respectively. Nitrate is also correlated with sulfate (0.548) and oxalate (0.611). Oxalate is



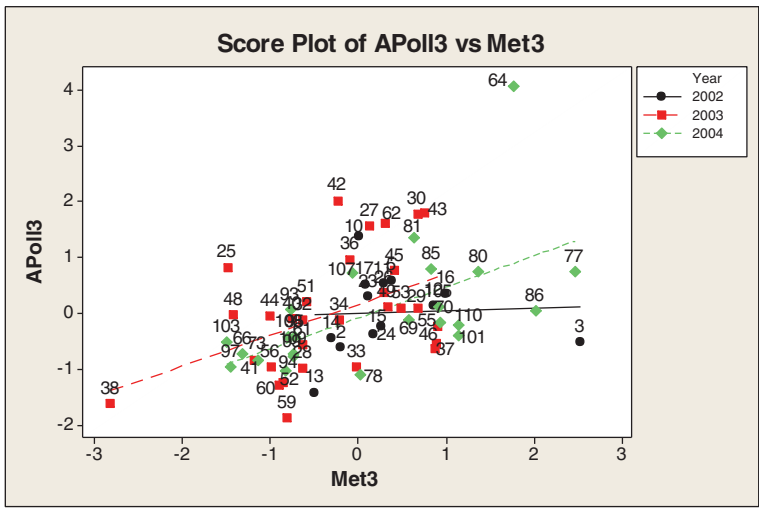




**Fig. 2** CCA score plots of three statistically significant canonical functions (A.Poll1 vs. Met1, A.Poll2 vs. Met2, and A.Poll3 vs. Met3). Scores are distinguished by year of collection (*different plotting symbols*). The number next to each point corresponds to the observation number, used to identify collection dates

**Table 3** Correlation matrix of air pollutants and meteorological parameters

| | A | C | N | S | Ox | $O_3$ | $NO_x$ | Sp | T | D |
|---|---|---|---|---|---|---|---|---|---|---|
| C | 0.392 | | | | | | | | | |
| N | 0.699 | 0.664 | | | | | | | | |
| S | 0.377 | 0.484 | 0.548 | | | | | | | |
| Ox | 0.520 | 0.458 | 0.611 | 0.752 | | | | | | |
| $O_3$ | 0.241 | 0.157 | 0.433 | 0.416 | 0.622 | | | | | |
| $NO_x$ | 0.302 | 0.365 | 0.357 | 0.026 | 0.073 | −0.424 | | | | |
| Sp | −0.127 | −0.058 | −0.228 | 0.010 | 0.043 | 0.165 | −0.379 | | | |
| T | 0.360 | 0.383 | 0.572 | 0.506 | 0.682 | 0.686 | 0.068 | 0.014 | | |
| D | 0.065 | 0.086 | 0.031 | −0.042 | −0.047 | −0.100 | 0.196 | 0.213 | 0.143 | |
| H | 0.078 | −0.163 | −0.176 | −0.212 | −0.292 | −0.376 | −0.117 | 0.074 | −0.448 | −0.127 |

*A* ammonium, *C* calcium, *N* nitrate, *S* sulfate, *Ox* oxalate, $O_3$ ozone, $NO_x$ nitrogen oxides, *Sp* wind speed, *T* temperature, *D* wind direction, *H* humidity

positively correlated with both ozone (0.622) and sulfate (0.752). This can explain why ozone was the only large canonical weight in A.Poll1, the first canonical variate for air pollutants. Numerical values of the simple correlations between original variables and their respective canonical variate are presented in Table 4.

Manly (2005) suggested interpreting loading values greater than ±0.5 and that threshold was implemented in this work. The correlations expressed by the loadings resulted in supplementary information to what was found in interpreting the canonical weights. Nitrate, sulfate, oxalate, and ozone are all positively correlated with A.Poll1 (Table 4). In the simple correlations of meteorological parameters, temperature is highly correlated with Met1 while humidity is negatively correlated with Met1. Therefore, A.Poll1 is a good measure of conditions of high ozone mixing ratios and oxalate concentration as well as moderate nitrate and sulfate concentrations; Met1 is a measure of high temperature and low humidity. Overall, this interpretation suggests higher temperature and lower humidity result in larger observed values of ozone and oxalate. This result agrees with the classification of these two pollutants as secondary, produced in the atmosphere due to various chemical reactions (Seinfeld and Pandis 1998). The correlation of nitrate and sulfate with temperature requires additional exploration.

**Table 4** Canonical loadings (structure correlations), correlations between original variables, and their canonical variates

| | A.Poll1 | A.Poll2 | A.Poll3 |
|---|---|---|---|
| Ammonium | 0.3372 | 0.3474 | 0.7874 |
| Calcium | 0.2842 | −0.0004 | 0.2895 |
| Nitrate | 0.5674 | −0.1178 | 0.5375 |
| Sulfate | 0.5347 | 0.1545 | −0.0120 |
| Oxalate | 0.8118 | 0.0655 | 0.2419 |
| Ozone | 0.9598 | 0.1257 | −0.0878 |
| $NO_x$ | −0.3679 | −0.6430 | 0.5826 |
| | Met1 | Met2 | Met3 |
| Wind speed | 0.2621 | 0.6229 | −0.5054 |
| Temp | 0.9439 | 0.0129 | 0.3105 |
| Wind direction | −0.1213 | 0.1948 | 0.5187 |
| Humidity | −0.5551 | 0.7359 | 0.2158 |

As ammonium nitrate's ($NH_4NO_3$) volatility increases with increasing temperatures and low humidity, reforming gaseous nitric acid ($HNO_3$) and ammonia in the troposphere (Du et al. 2010; Finlayson-Pitts and Pitts 2000), other species of particulate nitrate must contribute to the moderate, positive correlation between nitrate and temperature. Organonitrates have been measured in various aerosol studies in urban locations and were found to contribute significantly to organic mass of aerosols (Day et al. 2010; Garnes and Allen 2002; Liu et al. 2012). These molecules are present mainly due to the reaction between nitric oxide (NO) and peroxy ($RO_2$) radicals during the day (Liu et al. 2012; Finlayson-Pitts and Pitts 2000) as well as the product of alkene and nitrate radical reactions at night (Finlayson-Pitts and Pitts 2000). Based on this, organonitrates most likely play an important role in the atmospheric signature in Chicago with respect to nitrate formation.

Sulfate aerosols are produced in the atmosphere either via gaseous or aqueous phase oxidation of $SO_2$, sulfur dioxide. Both pathway reaction rates are dependent on a variety of factors, and, ultimately, oxidation of $SO_2$ in the aqueous phase is considered to be the major source of atmospheric sulfate aerosols (Finlayson-Pitts and Pitts 2000). This contradicts the canonical loading result of a moderate sulfate correlation with high temperature/low humidity. As Scheff et al. (1984) and Lee and Hopke (2006) found, the Midwest experiences significant amount of regional transport of sulfate; thus, locally measured sulfate may not have been produced in the immediate area. This would explain the contradiction in the CCA sulfate loading correlation result for A.Poll1 and Met1.

The only pollutant variable in A.Poll2 with a correlation value of ±0.5 is nitrogen oxides (−0.6430). Wind speed (0.6229) and humidity (0.7359) both have large positive correlations with Met2. These correlations confirm the results in the second linear combination for meteorology (Met2), which weighted humidity and wind speed significantly. However, in the linear combination of A.Poll2, ammonium was weighted highly positive, and nitrogen oxides, highly negative. Ammonium is not significantly correlated to A.Poll2 according to the canonical loading. Overall, this interpretation indicated that high winds are good predictors of low mixing ratios of nitrogen oxides.

Although the third canonical function was not interpreted due to its low canonical correlation, the third canonical loading was inspected for additional information. Original variables that correlated with A.Poll3 included ammonium, nitrate, and nitrogen oxides. Wind speed and the direction from which winds originated are moderately correlated with Met3. Therefore, when wind speed was lower and originating from a large degree (S, 180°; W, 270°; N, 0°/360°), higher ammonium and nitrate concentrations were measured along with high nitrogen oxide mixing ratios. This contradicts what was expressed earlier by the second pair of canonical weights regarding wind speed. Because the third canonical function is based on correlations not already expressed in the first two canonical variates and the canonical correlation was low (0.461), it is difficult to estimate the value of this canonical loading information. Thus, more consideration should be put on the wind speed result uncovered in the second canonical variate pair since its correlation was larger.

## Principal component analysis

Principal component analysis was applied to the original data described, using the correlation matrix as original data were measured on varying scale. Out of 11 PCs derived in PCA, the first four (PC1–PC4) were retained for further interpretation. These four principal components accounted for 77.0 % of the variability in the original data. The number of principal components retained was determined using Kaiser Criterion and a Scree plot. PC1 through PC4 each had eigenvalues greater than 1.0 and were retained. The Scree plot (Fig. 3) confirmed retention of PCs 1–3 for interpretation; however, it can be argued whether the slope between PC3 and PC4 is large enough to retain PC4 using the Scree plot. Table 5 displays each of the retained principal components, specifically the loading values corresponding to each of the original variables within the PCs. Only loading (absolute) values greater than or equal to ±0.3 were included in PC interpretations. The eigenvalue and variance explained by each PC are also displayed in Table 5, along with the cumulative percentage of variance of all of the PCs.

PC1 has the largest eigenvalue (4.0753) and explains 37.0 % of the original data variance. Ammonium (0.305), nitrate (0.383), sulfate (0.365), oxalate (0.431), and ozone (0.376) loading values in PC1 indicate a similar pattern with relation to temperature (0.402). The loading values are all positive, suggesting a moderately positive association between variables. All of these pollutants listed are secondary in nature, produced in the atmosphere due to chemical reactions (Seinfeld and Pandis 1998). Therefore, this PC is indicative of an overall measure of secondary aerosol production in the Chicago region. Because $NH_4^+$ is also weighted, PC1 could also convey neutralization of aerosol acidity as $NH_4^+$ neutralizes nitrate and sulfate.

PC2 explains 18.0 % of the total variance of the data, and its associated eigenvalue was 1.9784. PC2 identifies the consumption of $NO_x$ in the photochemical production pathway of tropospheric $O_3$ as their loadings are of opposite sign, −0.614 and 0.366, respectively. Additionally, wind speed (0.341) has a positive loading in PC2, while ammonium (−0.307) and calcium (−0.307) both have negative loadings. Note that the ammonium loading in this PC represents residuals that were not explained in PC1. This result shows that with large wind speeds, low concentrations

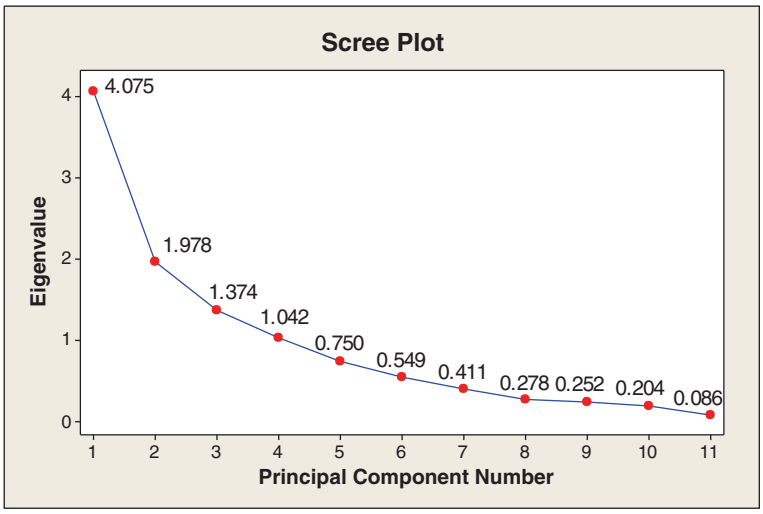



**Fig. 3** Scree plot used in determining the number of PCs to retain for interpretation

of these aerosol species are measured at the sampling site. Mixing of air masses due to wind may cause dilution of the species, resulting in low collection yields.

The only variables with significant loading values in PC3 were meteorological variables: wind direction (0.631), humidity (0.411), and wind speed (0.570). Wind speed residuals not explained by PC1 or PC2 were weighted in PC3. PC3 explains 12.5 % of the total variance in the original data. Due to variable weights, it can be inferred that PC3 is an overall measure of the state of meteorology during the sampling period. It is only possible to interpret how meteorology affects ammonium as the other pollutants' loading values are near zero. A correlation between humidity and $NH_4^+$ is present. Furthermore, the loading of wind direction (0.631) indicates that the direction from which the wind originates affects observed $NH_4^+$. Larger concentrations of $NH_4^+$ are present when wind direction is large, i.e., westerly direction. Up to this point, 67.5 % of the variance in the total data has been explained. PC4 explains an additional 9.5 % of the variance of the original data and is the last principal component retained for interpretation. Residuals of temperature (0.311), wind direction (0.525), and humidity (−0.635) not expressed by PCs 1–3 are expressed in PC4. The loading values for air pollutants were below the defined threshold for interpretation, thus PC4 does not yield any further information. The absence of new information in PC4 supports the findings from the Scree plot, which suggested retaining only PC1 through PC3 as the slope between PC3 and PC4 was small.

**Table 5** Principal components' loading values, eigenvalues, and percentage of variance explained by each PC. Cumulative variance of all PCs was 77.0 %

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| Ammonium | 0.305 | −0.307 | 0.251 | −0.146 |
| Calcium | 0.280 | −0.381 | 0.073 | −0.140 |
| Nitrate | 0.383 | −0.281 | −0.117 | −0.150 |
| Sulfate | 0.365 | −0.009 | 0.049 | −0.279 |
| Oxalate | 0.431 | 0.057 | −0.019 | −0.030 |
| $O_3$ | 0.376 | 0.366 | −0.112 | −0.013 |
| $NO_x$ | −0.066 | −0.614 | −0.094 | 0.286 |
| Wind speed | 0.102 | 0.341 | 0.570 | 0.023 |
| Temperature | 0.402 | 0.146 | −0.049 | 0.311 |
| Wind direction | 0.036 | −0.144 | 0.631 | 0.525 |
| Humidity | −0.211 | −0.097 | 0.411 | −0.635 |
| Eigenvalue | 4.0753 | 1.9784 | 1.3737 | 1.0424 |
| Variance (%) | 37.0 | 18.0 | 12.5 | 9.5 |
| Cum. var. (%) | 37.0 | 55.0 | 67.5 | 77.0 |

Score plots of PC2 vs. PC1 and PC3 vs. PC1, projecting original data in the dimensionality of principal components, are displayed in Fig. 4. The scores are differentiated by year of collection indicated by data points' shape and color (2002 round, 2003 square, and 2004 diamond). The scores on the positive end of PC1, circled and labeled no. 1, have a significant contribution from high temperature and large secondary air pollutants' recorded values, coinciding with the larger factor loadings on PC1 for these variables. For example, observations 23 and 46 were very warm (28.4 and 29.5 °C, respectively) with high ozone (78.3 and 97.9 ppb) and sulfate (18.8 and 22.7 μg/m$^3$)

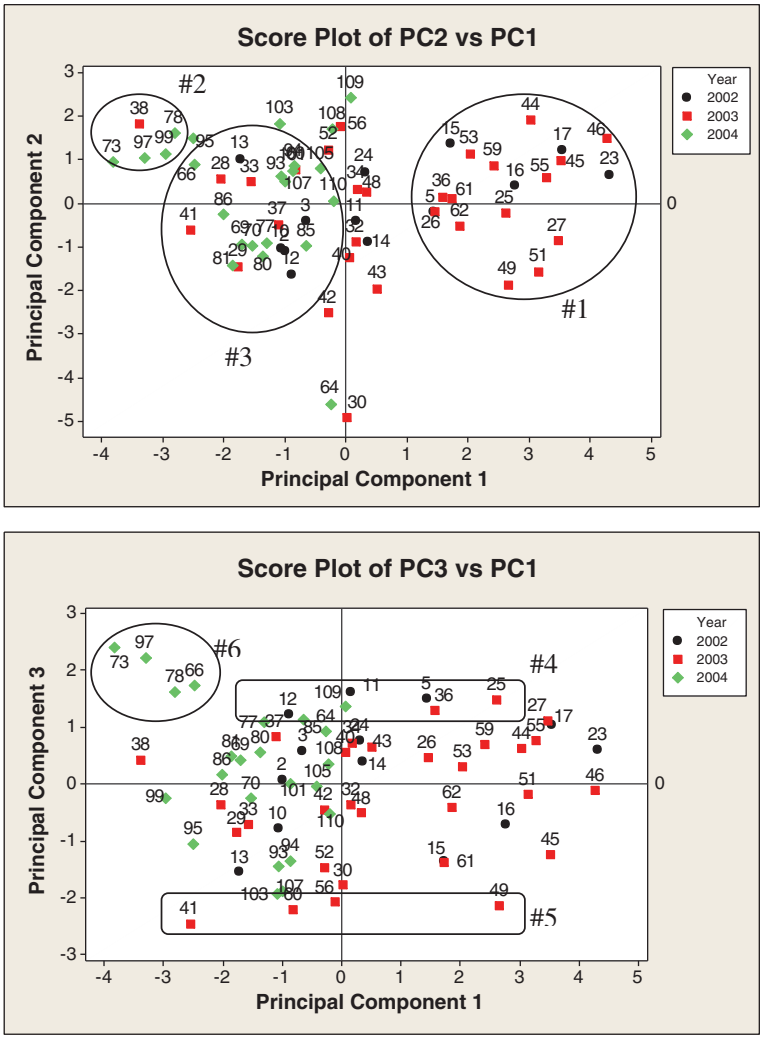



**Fig. 4** PCA score plots projecting original data in principal component space: **a** principal component 2 versus principal component 1, and **b** principal component 3 versus principal component 1. The label next to each point indicates the row/observation number, which was used to identify date of collection

concentrations. The majority of the points in the positive section of PC1 are from late morning/early afternoon collections which explain high temperatures and an increase in the oxidative nature of the atmosphere, resulting in large secondary air pollutant concentrations. On the contrary, observations in a small cluster labeled no. 2 at the far left, negative end of PC1, including 38, 73, 78, 97, and 99, were among the coolest days on record. For example, temperatures of 16.8 and 17.0 °C were recorded on observations 38 and 73, respectively. Many of the points are from morning aerosol collections, when temperatures are lower. The combination of low concentrations of nitrate, sulfate, oxalate, and ozone and high humidity may explain the negative scores on PC1 (circled nos. 2 and 3), as the loading value for humidity is moderately negative.

The difference between the positive and negative scores of points 30, 64, 56, 108, and 109 vertical axes in the PC2 vs. PC1 plot might be explained by $NO_x$ mixing ratios as well as $NH_4^+$, $O_3$, and wind speed data. Observations 30 and 64 had very large negative scores on PC2. $NO_x$ and $NO_3^-$ observed values were large while wind speed and $O_3$ were low for both points. Scores 56, 108, and 109 had very low $NO_x$ and $NO_3^-$ values and conversely had large $O_3$ and wind speed values. This explains the location of observations 56, 108, and 109 on the positive section of PC2. In summary, several groups of observations exhibiting similar pattern of meteorology and pollutant concentrations were identified in the score plot of PC2 vs. PC1.

In the interpretation of the PC3 vs. PC1 score plot the focus of interpretation was on PC3 (vertical axis),

as PC1 was explained earlier in the assessment of the first score plot. The position of the scores having positive or negative values on PC3 is highly dependent on the wind direction and wind speed during sampling. Scores on the positive side of PC3, such as 5, 10, 11, 25, and 36 circled no. 4 in the plot, are associated with high winds recorded from the northwest and southwest, moderate pollutant concentrations, and moderate/high humidity. Conversely, scores 30, 41, 49, 56, 60, 103, and 107 in the area labeled no. 5 have low recorded wind speeds originating from the north to the east and low pollutant concentrations. Humidity was moderate in both positive and negative score cases. The small cluster of points 66, 73, 78, and 97 circled and labeled no. 6 has several common recorded observations: high west/northwest winds and low pollutant concentrations. As meteorological parameters are weighted positive on PC3, this cluster is largely positively on PC3 axis. Scores near zero or about the origin are not particularly influenced by the large loading values on either PC. Overall, several groups of observations exhibiting similar pattern of meteorology and pollutants' concentrations were identified between the two score plots.

## Conclusions

By applying canonical correlation analysis to air pollutant (ammonium, calcium, nitrate, sulfate, oxalate, ozone, and nitrogen oxides) and meteorology (wind speed, temperature, wind direction, humidity) data, linear relationships were uncovered corresponding to the dependence of local pollutant concentration on meteorology. Three canonical functions derived in CCA were significant at the 0.05 level. Through interpretation of canonical weights, it was found that temperature influenced the mixing ratios of ozone positively. Canonical structure correlations also supported this and additionally revealed that oxalate was also positively correlated with temperature. Based on air studies in other urban cities, organonitrates may contribute to Chicago's atmospheric signature as a positive correlation was found between nitrate and temperature in the first canonical loading. A moderate correlation between sulfate and high temperature/low humidity in A.Poll1 and Met1 canonical structure correlations contradicts the literature as aqueous oxidation of $SO_2$ to form sulfate is a dominant pathway in the atmosphere. As previous studies in Midwestern cities found, regional transport might be a major component to sulfate aerosols in the Chicago area. Canonical weights of the second canonical function linked high concentrations of ammonium and nitrogen oxide mixing ratios to low wind speeds and high humidity. The corresponding canonical loading values confirmed the occurrence of high ammonium concentrations on humid days with large wind speed.

Principal component analysis was used to reduce original data dimensions from 11 to 4. The four derived principal components with eigenvalues greater than 1.0 captured 77.0 % of the variance in the original data. PC1 was an overall measure of local secondary air pollutants, whereas photochemistry and wind speed were expressed in PC2. PC3 was a measure of residual meteorological conditions not already expressed and suggested the influence of wind direction on $NH_4^+$ concentrations. PC4 did not present new information. Overall, using both multivariate statistical techniques resulted in independent and overlapping information about the relationships between air pollutant and meteorological variables in Chicago and also between air pollutant species themselves.

**Acknowledgments** The authors gratefully acknowledge and thank Sophonie Tchoua for his assistance with SAS® 9.2 software programming and coding in applying CCA to the dataset.

## References

Almeida, S. M., Pio, C. A., Freitas, M. C., Reis, M. A., & Trancoso, M. A. (2006). Source apportionment of atmospheric urban aerosol based on weekdays/weekend variability: evaluation of road re-suspended dust contribution. *Atmospheric Environment, 40*, 2058–2067.

Bascom, R., Bromberg, P. A., Costa, D. A., Devlin, R., Dockery, D. W., Frampton, M. W., et al. (1996). Health effects of outdoor air pollution. *American Journal of Respiratory and Critical Care Medicine, 153*, 3–50.

Braga, C. F., Teixeira, E. C., Meira, L., Wiegand, F., Yoneama, M. L., & Dias, J. F. (2005). Elemental composition of $PM_{10}$ and $PM_{2.5}$ in urban environment in South Brazil. *Atmospheric Environment, 39*, 1801–1815.

Day, D. A., Liu, S., Russel, L. M., & Ziemann, P. J. (2010). Organonitrate group concentrations in submicron particles with high nitrate and organic fractions in coastal southern California. *Atmospheric Environment, 44*, 1970–1979.

Du, H., Kong, L., Cheng, T., Chen, J., Yang, X., Zhang, R., et al. (2010). Insights into ammonium particle-to-gas conversion: non-sulfate ammonium coupling with nitrate and chloride. *Aerosol and Air Quality Research, 10*, 589–595.

Finlayson-Pitts, B. J., & Pitts, J. N., Jr. (2000). *Chemistry of the upper and lower atmosphere*. New York: Academic.

Fosco, T., & Schmeling, M. (2007). Determination of water-soluble atmospheric aerosols using ion chromatography. *Environmental Monitoring and Assessment, 130*, 187–199.

Fosco, T., & Schmeling, M. (2006). Aerosol ion concentration dependence on atmospheric conditions in Chicago. *Atmospheric Environment, 40*, 6638–6649.

Garnes, L. A., & Allen, D. T. (2002). Size distribution of organonitrates in ambient aerosol collected in Houston, Texas. *Aerosol Science and Technology, 36*, 983–992.

Google (2012). Google maps. maps.google.com. Accessed February 29, 2012.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1992). *Canonical correlation analysis. In Multivariate data analysis with readings* (pp. 193–222). New York: Macmillan Publishing Company.

Harrison, R. M., Smith, D. J. T., & Luhana, L. (1996). Source apportionment of atmospheric polycyclic aromatic hydrocarbons collected from an urban location in Birmingham, U.K. *Environmental Science and Technology, 30*, 825–832.

Hsieh, L.-Y., Chen, C.-L., Wan, M.-W., Tsai, C.-H., & Tsai, Y. I. (2008). Speciation and temporal characterization of dicarboxylic acids in $PM_{2.5}$ during a PM episode and a period of non-episodic pollution. *Atmospheric Environment, 42*, 6836–6850.

Johnson, R. A., & Wichern, D. W. (1998). *Canonical correlation analysis. In Applied multivariate statistical analysis* (pp. 587–627). Upper Saddle River: Prentice Hall.

Lee, J. H., & Hopke, P. K. (2006). Apportioning sources of $PM_{2.5}$ in St. Louis, MO using speciation trends network data. *Atmospheric Environment, 40*, S360–S377.

Lee, H. S., Wadden, R. A., & Scheff, P. A. (1993). Measurements and evaluation of acid air pollutants in Chicago using an annular denuder system. *Atmospheric Environment, 27A*, 543–553.

Liu, S., Shilling, J. E., Song, C., Hiranuma, N., Zaveri, R. A., & Russell, L. M. (2012). Hydrolysis of organonitrate functional groups in aerosol particles. *Aerosol Science and Technology, 46*, 1359–1369.

Manly, B. F. J. (2005). *Canonical correlation analysis. In Multivariate statistical methods: a primer* (pp. 143–161). New York: Chapman and; Hall/CRC.

Minitab Inc. (2010). Minitab 16 Statistical Software [Computer Software]. State College, PA.

Ravindra, K., Stranger, M., & Van Grieken, R. (2008). Chemical characterization and multivariate analysis of atmospheric $PM_{2.5}$ particles. *Journal of Atmospheric Chemistry, 59*, 199–218.

SAS Institute Inc. (2010). SAS 9.2 Statistical Software [Computer Software]. Cary, NC.

Scheff, P. A., Wadden, R. A., & Allen, R. J. (1984). Quantitative assessment of Chicago air pollution through analysis of covariance. *Atmospheric Environment, 18*, 1623–1631.

Schlesinger, R. B. (2007). The health impact of common inorganic components of fine particulate matter ($PM_{2.5}$) in ambient air: a critical review. *Inhalation Toxicology, 19*, 811–832.

Seinfeld, J. H., & Pandis, S. N. (1998). *Atmospheric chemistry and physics: from air pollution to climate change*. New York: Wiley.

Shen, Z., Cao, J., Arimoto, R., Han, Z., Zhang, R., Han, Y., et al. (2009). Ionic composition of TSP and $PM_{2.5}$ during dust storms and air pollution episodes at Xi'an, China. *Atmospheric Environment, 43*, 2911–2918.

Simcik, M. F., Eisenreich, S. J., & Lioy, P. J. (1999). Source apportionment and source/sink relationships of PAHs in the coastal atmosphere of Chicago and Lake Michigan. *Atmospheric Environment, 33*, 5071–5079.

Statheropoulos, M., Vassiliadis, N., & Pappa, A. (1998). Principal component and canonical correlation analysis for examining air pollution and meteorological data. *Atmospheric Environment, 32*, 1087–1095.

U.S. Census Bureau. (2010). 2010 Census interactive population search. http://2010.census.gov/2010census/popmap/ipmtext.php?fl=17. Accessed June 11, 2012.

U.S. Environmental Protection Agency. (2010). Our Nation's air: status and trends through 2008. http://www.epa.gov/airtrends/2010/report/fullreport.pdf. Accessed December 13, 2012.

Wilks, D. S. (2011). *Principal component (EOF) analysis. In Statistical methods in the atmospheric sciences* (pp. 519–562). New York: Elsevier Inc.

Yu, T. Y., & Chang, I. C. (2006). Spatiotemporal features of severe air pollution in northern Taiwan. *Environmental Science and Pollution Research, 13*, 268–275.

Zhou, F., Guo, H., & Liu, L. (2007). Quantitative identification and source apportionment of anthropogenic heavy metals in marine sediment of Hong Kong. *Environmental Geology, 53*, 295–305.