

Applying principal component pursuit to investigate the association between source-specific fine particulate matter and myocardial infarction hospitalizations in New York City

Rachel H. Tao^a, Lawrence G. Chillrud^b, Yanelli Nunez^{b,*}, Sebastian T. Rowland^b,
Amelia K. Boehme^a, Jingkai Yan^c, Jeff Goldsmith^d, John Wright^c, Marianthi-Anna Kioumourtzoglou^b

Background: The association between fine particulate matter (PM_{2.5}) and cardiovascular outcomes is well established. To evaluate whether source-specific PM_{2.5} is differentially associated with cardiovascular disease in New York City (NYC), we identified PM_{2.5} sources and examined the association between source-specific PM_{2.5} exposure and risk of hospitalization for myocardial infarction (MI).

Methods: We adapted principal component pursuit (PCP), a dimensionality-reduction technique previously used in computer vision, as a novel pattern recognition method for environmental mixtures to apportion speciated PM_{2.5} to its sources. We used data from the NY Department of Health Statewide Planning and Research Cooperative System of daily city-wide counts of MI admissions (2007–2015). We examined associations between same-day, lag 1, and lag 2 source-specific PM_{2.5} exposure and MI admissions in a time-series analysis, using a quasi-Poisson regression model adjusting for potential confounders.

Results: We identified four sources of PM_{2.5} pollution: crustal, salt, traffic, and regional and detected three single-species factors: cadmium, chromium, and barium. In adjusted models, we observed a 0.40% (95% confidence interval [CI]: –0.21, 1.01%) increase in MI admission rates per 1 µg/m³ increase in traffic PM_{2.5}, a 0.44% (95% CI: –0.04, 0.93%) increase per 1 µg/m³ increase in crustal PM_{2.5}, and a 1.34% (95% CI: –0.46, 3.17%) increase per 1 µg/m³ increase in chromium-related PM_{2.5}, on average.

Conclusions: In our NYC study, we identified traffic, crustal dust, and chromium PM_{2.5} as potentially relevant sources for cardiovascular disease. We also demonstrated the potential utility of PCP as a pattern recognition method for environmental mixtures.

Introduction

The association between exposure to air pollution and cardiovascular outcomes is well established, with increasing evidence

that fine particulate matter (PM_{2.5}) is particularly harmful.^{1–3} Inhalation of PM_{2.5} can lead to both chronic cardiovascular disease and acute events through several biologic pathways, including oxidative stress and inflammation, neural reflex arcs and autonomic imbalance, increased blood pressure, and translocation of inhaled PM_{2.5} constituents into systemic circulation.^{4,5} The relationship between PM_{2.5} exposure and acute cardiovascular events has been assessed primarily through time-series and case-crossover studies.^{3,6–8} A recent meta-analysis of 26 published studies found that each 10 µg/m³ increase in PM_{2.5} was associated with 1.02 times higher risk of myocardial infarction (MI).³

Several studies have explored the relationships between different sources of PM_{2.5} pollution and cardiovascular hospital admissions.^{9–11} The results of previous studies of cities on the East Coast of the United States suggest that PM_{2.5} from residual oil and traffic pollution may be associated with cardiovascular hospital admissions.^{10,11} A study examining source-specific PM_{2.5} and cardiovascular hospital admissions across New York State (NYS) also found traffic-related PM_{2.5} to be associated with MI.¹²

^aDepartment of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY; ^bDepartment of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY; ^cDepartment of Electrical Engineering, Columbia University, New York, NY; and ^dDepartment of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY.

This work was supported by grants NIEHS P30 ES009089, R01 ES028805, R01 ES030616 from the National Institute of Environmental Health Sciences.

All computing code used for this study can be publicly accessed via github.com/12taor/PCP-PM2.5-MI-paper. All otherwise publicly accessible data can also be found in the same GitHub repository. Health data from the Statewide Planning and Research Cooperative System (SPARCS) is not available as we do not have permission to share these data.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.enviroepidem.com).

*Corresponding Author. Address: Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032. E-mail: y.nunez@psehealthyenergy.org.

Copyright © 2023 The Authors. Published by Wolters Kluwer Health, Inc. on behalf of The Environmental Epidemiology. All rights reserved. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

Environmental Epidemiology (2023) 7:e243

Received: 9 August 2022; Accepted 20 January 2023

Published online 15 February 2023

DOI: 10.1097/EE9.000000000000243

What this study adds

In this study investigating the association between source-specific fine particulate matter (PM_{2.5}) and myocardial infarction hospitalizations in NYC, we demonstrate the potential utility of principal component pursuit (PCP) as a pattern recognition method for environmental mixtures research. PCP reduces the influence of extreme events on identification of consistent patterns, allowing for interpretable source apportionment of PM_{2.5} data with minimal subjective decision making by the researcher. We also provide an updated analysis of the association between specific sources of PM_{2.5} air pollution and cardiovascular disease focusing on NYC during the years 2007–2015.

To our knowledge, the most recent source-specific analysis of the association between $PM_{2.5}$ and cardiovascular disease specifically focusing on NYC was a 2011 paper,¹⁰ using 2001–2002 data. Since 2002, several policies aimed at improving air quality have been implemented on the federal, state, and local levels, including those aimed at reducing sulfur emissions from diesel fuel, as well as changes in regulations governing electricity generation. These changes and others may have changed the contributions of different pollution sources in NYC or altered the chemical composition of $PM_{2.5}$ coming from different air pollution sources.

Better understanding which sources of $PM_{2.5}$ pollution are most strongly associated with cardiovascular risk could aid in identifying targets for $PM_{2.5}$ reduction efforts. When assessing the effect of simultaneous exposure to multiple environmental exposures, researchers often use unsupervised dimensionality-reduction methods, such as Positive Matrix Factorization (PMF), Principal Components Analysis (PCA), or other factor analytic approaches, to identify patterns representing underlying pollution sources. When used with environmental mixtures data, these dimensionality-reduction approaches can be susceptible to outlying and extreme events, leading to reduced ability to find an interpretable solution.

We applied principal component pursuit (PCP) as a pattern recognition method to identify sources of $PM_{2.5}$ pollution using speciated $PM_{2.5}$ data from three NYC locations. PCP is advantageous for use with environmental mixtures data, as it is more robust to extreme events than other commonly used dimensionality-reduction methods. We examined associations between the identified $PM_{2.5}$ sources and MI admissions in NYC from 2007 to 2015, leveraging the New York Department of Health Statewide Planning and Research Cooperative System (SPARCS) database.

Methods

Study population

Daily hospitalization data for MI in NYC were extracted from the SPARCS database¹³ from 2007 to 2015. We used daily city-wide counts of acute-care admissions for MI as the dependent variable in the health models. The study population consists of people who received care for MI in acute-care facilities in NYS and resided in NYC.¹³

Outcome Assessment

MI admissions were identified based on International Classification of Diseases 9th Revision (ICD-9) for years before 2015 and based on International Classification of Diseases 10th Revision (ICD-10) for 2015. Admissions were identified as cases if ICD-9 code 410.x1 or ICD-10 code I21 occupied one of the first four diagnostic positions. We excluded “childbirth” or “trauma” admission types. We excluded MI readmissions that took place within 2 days of a previous MI admission, for a final sample size of 444,295 MI admissions.

Exposure Assessment

Ambient $PM_{2.5}$ concentrations and its constituents were extracted from the Air Quality System (AQS) database maintained by the United States Environmental Protection Agency (EPA).¹⁴ Samples for speciated $PM_{2.5}$ are collected every third or sixth day, and concentrations are reported in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$). The EPA measures total $PM_{2.5}$ concentration using gravimetric analysis and chemical constituents of $PM_{2.5}$ using a variety of methods for different chemicals, including X-Ray Fluorescence, ion chromatography, and thermal optical transmittance analysis. AQS places limitations on acceptable values for all measurements, which reflect theoretical limits of the measurement plus or minus some degree of uncertainty.¹⁴

We used data from three locations in NYC: lower Manhattan, southern Bronx, and northwestern Queens (Figure S1; <http://links.lww.com/EE/A217>). Data collection from the monitors used in this study occurred on the same schedule, such that scheduled date of data collection for all monitors coincided every 3 days. Individual monitors had lapses in data collection, so data were not available for all monitors on each observation day. Data from two separate monitors in lower Manhattan were used because they were close in space and did not overlap temporally. We used data from these monitor locations to compute the daily average concentrations on each of 978 total observation days for total $PM_{2.5}$ and 26 $PM_{2.5}$ constituents across NYC using information from any of the monitors with available data. We used average daily concentrations of total $PM_{2.5}$ and the following constituents: aluminum (Al), ammonium (NH_4), arsenic (As), barium (Ba), bromine (Br), cadmium (Cd), calcium (Ca), chlorine (Cl), elemental carbon (EC), organic carbon (OC), chromium (Cr), copper (Cu), iron (Fe), lead (Pb), magnesium (Mg), manganese (Mn), nickel (Ni), potassium (K), selenium (Se), silicon (Si), sodium (Na), sulfur (S), titanium (Ti), nitrate (NO_3), vanadium (V), and zinc (Zn).

Covariates

Temperature, pressure, and specific humidity data were extracted from the National American Land Data Assimilation System, NLDAS-2 Forcing.¹⁵ NLDAS reports hourly parameter values for 0.125° grids ($\sim 11 \text{ km} \times 14 \text{ km}$ in NYS). We took the average of the 24 hours provided in the NLDAS dataset for each day and aggregated daily values for the 0.125° grids to the geographic extent of NYC via population-weighted averaging at the census tract level.

Statistical analysis

Source apportionment

PCP is a dimensionality-reduction method used primarily in computer vision and signal processing applications, that can be understood as a robust form of PCA.¹⁶ We have adapted PCP to be used as an exposure pattern recognition method for environmental mixtures.¹⁷ PCP decomposes the exposure matrix into: (1) a low-rank matrix containing consistent patterns in the mixture and (2) a sparse matrix containing unique or extreme exposure events. By separating extreme exposure events from consistent patterns of exposure, PCP reduces the influence of extreme events on identification of consistent patterns. This approach is therefore more robust to extreme events than other pattern recognition methods, such as PCA or factor analysis alone. In a recent study using simulated environmental exposure data, PCP overall outperformed PCA in most simulated scenarios; however, when noise in the data was high, PCA and PCP performed similarly.¹⁷ When used in computer vision applications, PCP has the additional advantage of reducing researcher subjectivity by using theoretically-optimal, single, universal regularization parameters to generate the low-rank matrix.^{16,18} However, environmental data are particularly noisy compared with other PCP applications; we found that the default PCP regularization parameters resulted in overly low-rank matrices (i.e., ranks 1 or 2 were preferred; but we would expect a larger number of sources), and instead applied cross-validation to select the hyperparameters.

We used square-root PCP ($\sqrt{\text{PCP}}$), an extension of PCP,¹⁸ and combined it with a separate extension introducing a nonconvex penalty on the low-rank matrix.¹⁷ We used cross-validation to select the optimal rank of the low-rank matrix, which can be understood as the number of underlying patterns in the $PM_{2.5}$ data. Please see the Supplement for further details on hyperparameter selection. As in Gibson et al.,¹⁷ we used a version of the algorithm modified to allow for missing values, to make it possible

to include dates with missing measurements of some PM_{2.5} constituents, and we constrained the low-rank matrix to be nonnegative.

We subsequently applied nonnegative matrix factorization (NMF) to the PCP-generated low-rank matrix, to extract chemical loadings and factor scores for each observation day. We will refer to these chemical loadings and factor scores as PCP-NMF loadings and PCP-NMF scores. We identified pollution sources from examination of PCP-NMF loadings using a combination of expert knowledge and prior literature. Once pollution sources were identified from PCP-NMF loadings, we estimated daily concentrations of source-specific PM_{2.5}. We first regressed daily total PM_{2.5} concentration on daily PCP-NMF scores for each identified source, including a term for the daily sum of scores from the sparse matrix to account for PM_{2.5} not explained by the identified sources. We then multiplied each PCP-NMF score by the regression coefficient of its source in the model to compute daily source-specific PM_{2.5} concentrations.

Time-series health models

Once PM_{2.5} sources were identified, we conducted a time-series analysis using a Poisson regression model and quasi-likelihood to account for potential outcome overdispersion. To determine which sources were associated with MI as same-day exposures, estimated source-specific PM_{2.5} concentrations for each of the identified sources were simultaneously included in the regression model as predictors. We controlled for the following covariates as potential confounders: nonlinear terms (natural splines) for same-day (lag 0) ambient temperature (degrees of freedom, df = 4), 3-day average (average lag 1–3) ambient temperature (df = 4), same-day relative humidity (df = 4), and 3-day average relative humidity (df = 3), day of week indicators, and a natural spline term with 36 df (= 4 seasons × 9 years) to account for seasonal and long-term trends. To test for deviations from linearity, we ran separate generalized additive models for each PCP-NMF

factor, where the factor in question was modeled using a penalized spline while controlling for linear terms of all other sources, along with all covariates. Quasi Akaike's Information Criterion (qAIC) was used to determine whether each PCP-NMF factor should be modeled linearly or nonlinearly in the final model.

Effect estimates are presented as percent changes in MI hospital admission rates per 1 µg/m³ increase in source-specific PM_{2.5}, if the associations were linear. If nonlinear, we present the full exposure-response curve.

Statistical analyses were conducted using R version 4.0.2 (2020-06-22),¹⁹ the *pcpr*, *NMF*,²⁰ and *mgcv*²¹ packages.

Sensitivity analyses

In addition to same-day exposure, we also assessed the potential association between source-specific PM_{2.5} and MI admission rate at lags 1 and 2. In a second sensitivity analysis, we ran single-source health models. Finally, to examine the extent to which the health effect estimates may be driven by outliers in the PCP-NMF scores, we removed values more than 3 standard deviations away from the mean of each source and repeated analyses as described above. Two hundred fifty-eight observations were removed in the sensitivity analysis.

Results

During 2007–2015, the daily mean total PM_{2.5} concentration was 10.3 µg/m³ (SD: 6.0), and the daily median number of admissions for MI was 135 (IQR: 27) over the 978 observation days included in our analysis. MI admissions decreased overall from 2013 to 2015 and followed a seasonal pattern with highest MI rates in the winter (Figure S2, <http://links.lww.com/EE/A217>). Average ambient temperature was 12.3°C (SD: 9.7) (Table 1). Pearson correlation coefficients among PM_{2.5}

TABLE 1. Summary Statistics for Daily Number of ED Visits for Myocardial Infarction in NYC (2007–2015), Total PM_{2.5}, PM_{2.5} Constituents, Daily Ambient Temperature and Relative Humidity

	Minimum	25th%ile	Median	75th%ile	Maximum	Mean	SD
Total PM _{2.5} (µg/m ³)	1.6	5.9	8.9	13.0	38.6	10.3	6.0
Aluminum (Al)	0.0	7.4	17.0	29.3	309.7	22.5	25.4
Ammonium (NH ₄)	0.0	403.1	815.0	1,435.8	7,260.0	1,119.3	1,040.8
Arsenic (As)	0.0	0.0	0.3	0.8	4.0	0.5	0.6
Barium (Ba)	0.0	0.0	0.0	2.2	52.6	1.9	4.3
Bromine (Br)	0.0	1.7	2.5	4.0	58.5	3.0	2.6
Cadmium (Cd)	0.0	0.0	0.0	2.3	23.0	1.7	3.0
Calcium (Ca)	0.0	30.3	45.5	64.4	765.4	51.7	36.9
Chlorine (Cl)	0.0	4.0	10.0	26.0	1,550.0	36.8	89.0
Chromium (Cr)	0.0	0.3	1.0	2.0	144.8	2.1	7.5
Copper (Cu)	0.0	2.6	3.9	5.7	39.5	4.6	3.4
Elemental Carbon (EC)	80.5	434.9	604.9	823.6	6,170.0	706.9	463.7
Iron (Fe)	9.8	69.6	95.0	130.7	552.8	105.5	53.1
Lead (Pb)	0.0	0.5	1.5	2.7	45.7	2.0	2.4
Magnesium (Mg)	0.0	0.0	3.6	9.3	129.0	7.3	11.7
Manganese (Mn)	0.0	1.0	1.7	2.8	12.9	2.1	1.7
Nickel (Ni)	0.0	1.9	3.5	6.3	44.5	4.9	4.6
Organic Carbon (OC)	566.0	1,775.0	2,465.8	3,348.8	9,990.0	2,706.6	1,270.0
Potassium (K)	0.0	10.0	29.4	52.0	909.7	39.3	58.3
Selenium (Se)	0.0	0.0	0.2	0.5	4.2	0.4	0.6
Silicon (Si)	1.0	32.5	49.1	72.9	686.0	61.3	53.4
Sodium (Na)	0.0	31.5	67.9	125.0	1,185.0	95.1	102.9
Sulfur (S)	47.4	395.0	628.0	986.8	4,828.4	791.3	624.2
Titanium (Ti)	0.0	1.0	2.0	3.3	20.7	2.4	2.2
Nitrate (NO ₃)	97.0	530.1	1,002.3	2,113.1	11,700.0	1,604.1	1,581.9
Vanadium (V)	0.0	0.6	1.7	4.0	26.2	2.9	3.3
Zinc (Zn)	0.0	11.5	19.8	33.4	346.2	25.9	23.0
Temperature (°C)	-14.0	4.3	12.7	21.2	31.5	12.3	9.7
Relative Humidity (%)	0.4	0.7	0.8	0.8	1.0	0.7	0.1
MI Counts	10.0	121.0	135.0	148.0	216.0	135.2	196.6

All values without marked units are PM_{2.5} constituents and are reported in ng/m³.

constituents ranged from -0.1 to 0.9, with highest correlations between ammonium and sulfur, ammonium and nitrate, aluminum and silicon, chlorine and sodium, magnesium and sodium, manganese and nickel, and OC and sulfur (Figure S3, <http://links.lww.com/EE/A217>).

Source apportionment

Using cross-validation, we estimated rank $r = 7$ for the low-rank matrix, which we used as the number of expected sources in NMF. Four sources of $PM_{2.5}$ and three single-constituent factors were identified based on PCP-NMF loadings: (1) crustal dust, (2) salt, (3) traffic, (4) regional, (5) cadmium, (6) chromium, and (7) barium. Regional, crustal, and traffic $PM_{2.5}$ contributed the largest approximate proportion of variance (Figure S4, <http://links.lww.com/EE/A217>).

We examined the PCP-NMF loadings, along with seasonal, long-term, and weekly patterns in estimated $PM_{2.5}$ for

each identified $PM_{2.5}$ source (Figures 1; Figures S5 and S6; <http://links.lww.com/EE/A217>). The crustal dust source was characterized by high levels of silicon, aluminum, and titanium. It was highest in the summer, lowest in the winter, and was higher on weekdays than weekends. Salt was primarily composed of chlorine, sodium, and magnesium. It did not follow a weekly pattern but appeared to be highest in the spring, with occasional autumnal and winter peaks. Traffic was characterized by high levels of zinc, nickel, nitrate, elemental carbon, calcium, copper, lead, iron, manganese, and vanadium. It was highest in winter and on weekdays and decreased slightly during the study period. The regional source was characterized by high loadings for sulfate, ammonium, organic carbon, and nitrate, along with selenium and potassium. It had both summer and winter peaks most years, decreased over the study period, and was slightly higher on weekends than weekdays (Tables S1 and S2; <http://links.lww.com/EE/A217>).

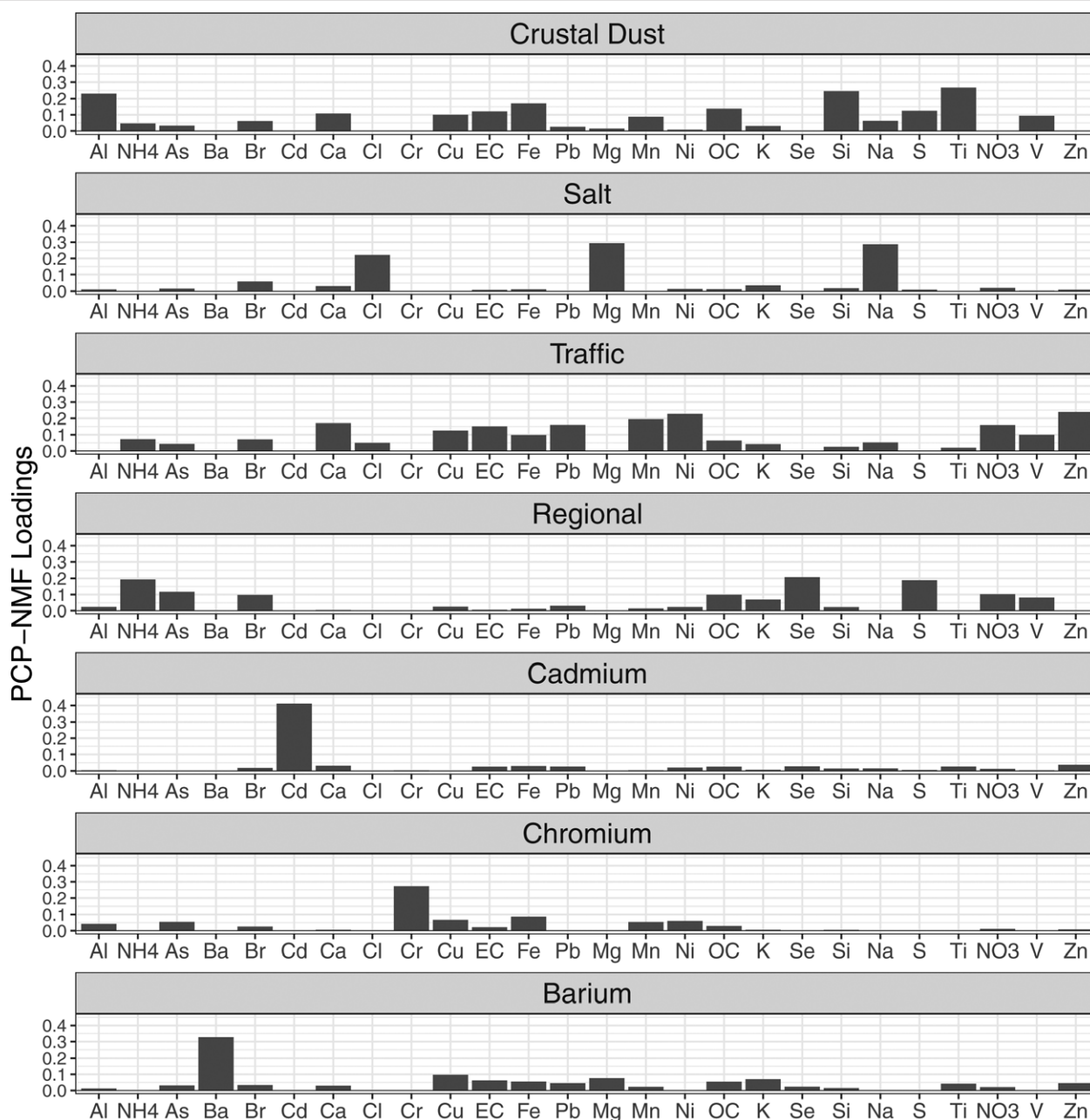


Figure 1. PCP-NMF loadings for chemical constituents of $PM_{2.5}$. Constituents are listed using chemical formulas or abbreviations.

Downloaded from <http://journals.lww.com/environepidem> by BHDMSFPHKAVI7Eoum1QIN4a+kJLHEZgbsHh04XMI0
 NCjwCX1AMN YqP/llQHID3j3D00ORy7TVSFI4C3V/C1y0abggQZxdgJ2MwZLe= on 03/31/2024

The latter three factors—cadmium, chromium, and barium—were each predominantly characterized by a single $PM_{2.5}$ constituent (Figure 1). The cadmium factor did not follow a seasonal or weekday pattern. The chromium factor appeared to be higher on Fridays than other days of the week but did not differ between weekday and weekend days and had two peaks over the study period, in autumn 2009 and 2013. The barium factor did not have weekday or seasonal trends but appeared to peak at the end of 2015 (Tables S1 and S2; <http://links.lww.com/EE/A217>).

PCP-NMF scores for each factor over time were not strongly correlated with one another—the highest Pearson correlation between two factors was 0.4, between traffic and the secondary sulfate (Figure S7; <http://links.lww.com/EE/A217>).

Sparse matrix

Three point six percent (3.6%) of the sparse matrix was populated with nonzero elements, representing extreme events that could not be explained by the consistent patterns identified in the low-rank matrix. Some sparse events were detected for all chemical constituents of $PM_{2.5}$. Notably, we observed sparse events for potassium in early July—indicating fireworks—for most years and in late December or early January—for New Year's celebrations—for several years (Figure S8; <http://links.lww.com/EE/A217>).

Time-series health analysis

For all factors except salt and cadmium, we detected no deviations from linearity based on qAIC. We observed a 0.40% (95% CI: -0.21, 1.01%) increase in MI rates per 1 $\mu\text{g}/\text{m}^3$ increase in same-day traffic $PM_{2.5}$, a 0.44% (95% CI: -0.04, 0.93%) increase in MI rates per 1 $\mu\text{g}/\text{m}^3$ increase in same-day crustal $PM_{2.5}$, and a 1.34% (95% CI: -0.46, 3.17%) increase in MI rates per 1 $\mu\text{g}/\text{m}^3$ increase in same-day chromium-related $PM_{2.5}$, on average, adjusting for confounders (Figure 2). For all other factors, the association was null (Table S3; <http://links.lww.com/EE/A217>; Figure 2). We present the estimated nonlinear curves for salt and cadmium in Figures S9 and S10; <http://links.lww.com/EE/A217>.

At lags 1 and 2 with full data, the effect estimates were closer to null than at lag 0 for most sources, including traffic, crustal,

and chromium $PM_{2.5}$. We observed a 0.14% (95% CI: -0.04, 0.33%) increase in MI rates per 1 $\mu\text{g}/\text{m}^3$ increase in regional $PM_{2.5}$ at lag 1 and a 0.13% (95% CI: -0.05, 0.31%) increase at lag 2, whereas the effect was null at lag 0. When outliers were removed, we continued to observe a weak positive association between regional $PM_{2.5}$ and MI at lag 2 (0.23%; 95% CI: -0.06, 0.53%) but at lags 0 and 1 the association was null. We observed a 0.59% (95% CI: -0.27, 1.45%) increase in MI rates per 1 $\mu\text{g}/\text{m}^3$ increase in barium $PM_{2.5}$ at lag 2, whereas the effect was null at lags 0 and 1. When outliers were removed, we observed a null association between barium and MI admission rate at all 3 lags. We observed a 1.01% (95% CI: 0.17, 1.85%) increase in MI admission rate per 1 $\mu\text{g}/\text{m}^3$ increase in crustal $PM_{2.5}$ at lag 1 with outliers removed (Figure S11; <http://links.lww.com/EE/A217>; Tables S3 and S4; <http://links.lww.com/EE/A217>).

In single-source models (using full data), a 1 $\mu\text{g}/\text{m}^3$ increase in crustal dust was associated with a 0.54% (95% CI: 0.09, 1.00%) increase in MI admission rate and a 1 $\mu\text{g}/\text{m}^3$ increase in traffic was associated with a 0.59% (95% CI: 0.06, 1.12%) increase in MI admission rate, adjusting for all covariates. The observed association between same-day chromium and MI admission rate in the full model was attenuated in the single-source model, where a 1 $\mu\text{g}/\text{m}^3$ increase in chromium was associated with a 1.16% (95% CI: -0.62, 2.98%) in the single-source model (Table S5; <http://links.lww.com/EE/A217>).

After removing outlying scores larger than 3 SD from the mean score for each source, percent change in MI rate per 1 $\mu\text{g}/\text{m}^3$ increase in same-day traffic-related $PM_{2.5}$ increased to 1.02% (95% CI: 0.04, 2.00%). Percent change in MI rate per 1 $\mu\text{g}/\text{m}^3$ increase in same-day crustal dust and chromium-related $PM_{2.5}$ decreased to 0.11% (95% CI: -0.73, 0.95%) and -1.01% (95% CI: -6.94, 4.28%), respectively. Salt, which had a null association with MI in the main analysis, appeared to have a negative relationship with MI when outliers were removed (-2.63%; 95% CI: -5.60, 0.43%) (Figure S12; <http://links.lww.com/EE/A217>; Table S4; <http://links.lww.com/EE/A217>).

Discussion

Using data from EPA's publicly available AQS database and a robust exposure pattern recognition method, we identified four sources of $PM_{2.5}$ and three single-constituent factors in

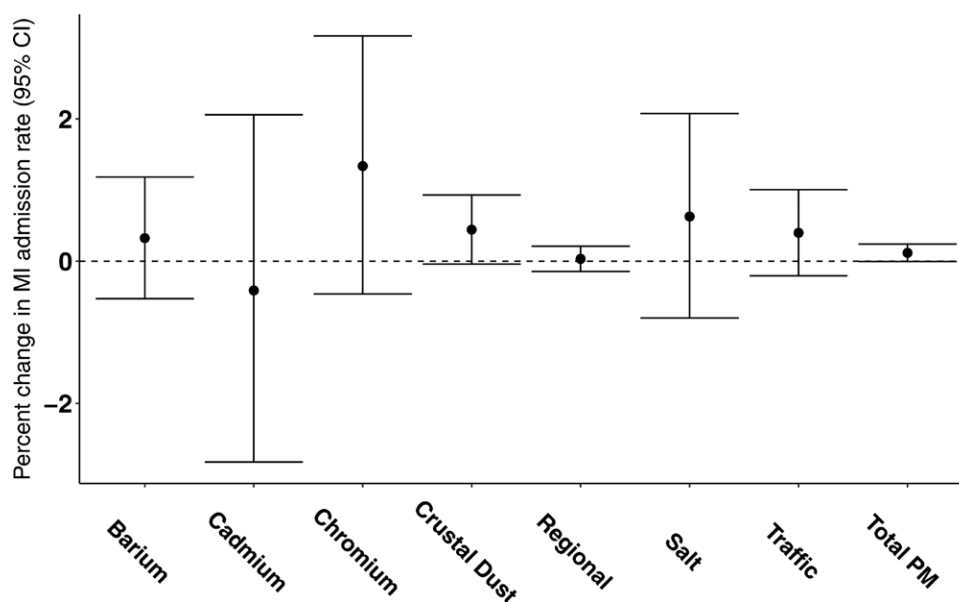


Figure 2. Forest plot of percent change in MI admission rates per 1 $\mu\text{g}/\text{m}^3$ increase in source-specific $PM_{2.5}$, adjusting for same-day and 3-day average temperature, same-day and 3-day average relative humidity, day of the week, and seasonal and long-term trends.

NYC between 2007 and 2015: (1) crustal dust, (2) salt, (3) traffic, (4) regional, (5) cadmium, (6) chromium, and (7) barium. Leveraging data from SPARCS, we observed increased rates of MI admissions with increased traffic, crustal dust, and chromium $PM_{2.5}$, but not for same-day salt, regional, cadmium, or barium. We observed marginal associations between lag 1 and 2 regional $PM_{2.5}$ and increased MI admission rates. After removing outliers, we continued to observe increased rates of MI admission with increased traffic-related and regional $PM_{2.5}$.

Source apportionment

Crustal dust

Crustal dust, containing high levels of silicon, aluminum, and titanium, can come from natural sources such as soil and can also mix with suspended road dust and construction dust.^{10,22,23} Our results are consistent with other studies, which also found higher concentrations of crustal dust during the summer months.²²

Salt

The salt source had high levels of chlorine, sodium, and magnesium, and appeared to increase in the spring and decrease in the autumn. Although some source-apportionment studies have identified a similar salt source,^{11,24} others have identified more than one salt component, differentiating between fresh and aged sea salt originating from different parts of the United States.^{22,23}

Traffic

The traffic source consisted of elemental carbon, nitrate, ammonium, zinc, copper, iron, and lead. $PM_{2.5}$ from traffic exhaust consists of high levels of elemental and organic carbon, as well as ammonium nitrate, iron, copper, and zinc.^{22,23,25} The presence of lead could indicate that particles from road dust also load on this factor, as lead, copper, iron, and zinc are commonly used in brake lining materials, which contribute to road dust.²⁶ Previous studies in NYC and other East Coast cities have either identified separate patterns for $PM_{2.5}$ from tailpipe emissions and $PM_{2.5}$ from road dust^{11,23} or a single component representing a mixture of emissions exhaust, resuspended road dust, and tire/brake wear.¹⁰

$PM_{2.5}$ concentrations from traffic decreased slightly over the course of the study period. Since nitrate is an important constituent of the traffic factor, it is possible that this trend was related to reductions in nitrate traffic emissions from policy changes that occurred during this time. Relevant policy changes include the Tier II Tailpipe NOx Emissions Standard for light-duty vehicles, implemented between 2004 and 2010, and the Clean Heavy-Duty Bus and Truck Rule, which requires all new heavy-duty diesel vehicles sold after July 1, 2007, to have particle control traps and those sold after January 1, 2010, to have NO_x controls.²⁷

Regional

The regional source was characterized by high levels of sulfate, ammonium, nitrate, and organic carbon. From 2007 to 2009, this source had a decreasing overall trend with clear summer peaks, plateauing after 2009. Sulfate in NYC typically originates from sulfur dioxide emissions by coal-fired power plants in the upper Ohio River Valley.^{28,29} Our results are consistent with other recent $PM_{2.5}$ source-apportionment analyses in NYS, and the overall decrease in regional $PM_{2.5}$ concentrations since 2007 is likely attributable to decreased use of coal for power generation.²³

Single-constituent factors: cadmium, chromium, and barium

We detected three single-species $PM_{2.5}$ sources: cadmium, chromium, and barium. These sources may be related to industrial emissions originating from chemical and metal processing, coke production, and metal recycling in NYS.²³

The chromium and barium factors both appeared to have extreme events during the study period, which we might have expected to be separated into the sparse matrix instead. This result demonstrates that extreme events may appear in the low-rank matrix if they are consistent with factors detected in the low-rank matrix.

Health models

Traffic

We observed that same-day traffic-related $PM_{2.5}$ was associated with an increase in MI admission rates. Traffic-related $PM_{2.5}$ has been associated with cardiovascular disease in prior studies, and several common constituents of traffic emissions are known to be associated with systemic inflammation.^{10–12,30–32} A source apportionment and health analysis using PMF in NYS detected an increase in MI admissions rates per IQR increase in spark-ignition emissions but a null association with diesel emissions.¹²

The observed association between traffic and MI was detected in both full and single-source models and remained robust after removing outliers. The effect estimate increased when outliers were removed, suggesting that the outliers that were removed were driving the association downward in the main analysis. Since the traffic factor peaked in winter months, this result could be attributable to exposure measurement error: traffic-related $PM_{2.5}$ levels are highest in the winter, when most people are likely to keep their windows closed and spend more time indoors, decreasing their exposure to outdoor traffic-related $PM_{2.5}$.

Crustal dust

We observed an increase in MI admission rates associated with crustal $PM_{2.5}$, predominantly composed of silicon and aluminum. Silicon as a chemical constituent of $PM_{2.5}$ has been linked to cardiovascular mortality,^{33,34} and mortality related to $PM_{2.5}$ has been found to be modified by increased proportion of aluminum.³⁵ Aluminum and silicon as $PM_{2.5}$ constituents have been linked to inflammation and oxidative stress.³⁶ Prior literature on the potential association between crustal $PM_{2.5}$ and cardiovascular disease has been mixed, with some studies reporting strong and others null associations.^{10–12,23}

Although the observed association between same-day crustal dust and MI was detected in both full and single-source models, when outliers were removed, the association became null, suggesting that the apparent association detected in the main analysis may have been driven by outliers. In contrast, at lag 1, we observed a positive association between crustal $PM_{2.5}$ and MI admissions with outliers removed, which had not been detected at lag 1 with full data. This result suggests that the outliers that were removed could have been driving the association toward the null at lag 1.

Chromium

We observed an increase in MI admission rates associated with same-day chromium $PM_{2.5}$ in the full model but not the single-source model. Though chromium is not generally considered to be associated with MI,³⁷ chromium air pollution was found to be associated with increased risk for cardiovascular disease in a study based in Xi'an, China.³⁸ When extreme outliers were removed, the observed association between the chromium

source and MI admissions became null, suggesting that the apparent association observed in the main analysis may have been driven by outliers.

Overall

For all three sources where we observed a positive association with MI admission rate, the effect estimate was either comparable to or greater than that of total $PM_{2.5}$ on a per $1 \mu\text{g}/\text{m}^3$ basis. The point estimates for traffic and crustal dust were lower than that of chromium (as well as salt, which appeared to have a null association with MI), indicating that chromium $PM_{2.5}$ is likely more toxic for MI. It should be noted, nonetheless, that the estimated concentrations of traffic and crustal $PM_{2.5}$ are higher than that of chromium $PM_{2.5}$ and, therefore, the expected overall impact on MI cases is also expected to be larger. The overall decrease that we observed in traffic $PM_{2.5}$ over the study period, combined with its high toxicity on a per $\mu\text{g}/\text{m}^3$ basis, suggest that traffic-related policy changes implemented between 2004 and 2010, such as the Tier II Tailpipe NOx Emissions Standard and the Clean Heavy-Duty Bus and Truck Rule, may have had a positive impact on MI hospital admissions since implementation.

Strengths

With PCP, we were able to automatically remove extreme exposure events from the low-rank matrix into the sparse matrix, rendering the resulting factor analysis of the low-rank matrix more interpretable. An important example of the utility of the sparse matrix is annual fireworks events. Fireworks produce high concentrations of potassium ion, but usually only produce a discernible signal on festival days, such as 4th of July (in the United States) and New Years' Eve. Other dimensionality-reduction methods, such as PCA and factor analysis alone, require researchers to manually remove observations on the days surrounding these holidays, as these annual extreme exposure events can make the solutions difficult to interpret.^{10,24} This process places the burden on the researcher to decide which extreme observations to remove and how to determine criteria for removal. Using PCP, observations that are not consistent with the patterns within the low-rank matrix are automatically separated into the sparse matrix. In our analysis, we found that most of the sparse events for potassium occurred on or near 4th of July or New Years' Eve, but sparse potassium events did not occur on these dates every year, and some sparse potassium events occurred on other dates during the year.

Our study had several other strengths, including the leveraging of the SPARCS dataset and EPA's AQS database, both of which allowed for a long study. The public availability of the AQS database also improves reproducibility of our source-apportionment analysis, which is available on GitHub.

Limitations

Through this analysis, we identified a few limitations to PCP as a source-apportionment method for air pollution research. As with other dimensionality-reduction techniques, a fully interpretable solution for PCP is not guaranteed, and we generated three factors that essentially comprised a single $PM_{2.5}$ constituent without being identifiable as specific pollution sources. Some of the sources identified, such as chromium, explain a small proportion of total estimated $PM_{2.5}$ and results for those sources should be interpreted with caution. Additionally, it is not possible to directly compare the relative variance in total $PM_{2.5}$ explained by the sparse versus low-rank matrix once separated by PCP. Although one of PCP's advantages is that it separates extreme

exposure events into the sparse matrix, we found that this separation does not necessarily preclude the existence of extreme events within the low-rank matrix, if these outlying events are consistent with the long-term pattern detected in the low-rank matrix. We found extreme events in both the cadmium and the barium sources within the low-rank matrix; we removed these potentially outlying events from the health models in sensitivity analyses. Finally, we found that when applying nonconvex \sqrt{PCP} to speciated $PM_{2.5}$ data, it was necessary to tune hyperparameters, which is a time-intensive process that comes with a certain degree of researcher subjectivity. Other formulations of PCP have used theoretically-optimal single universal values for hyperparameters λ and μ ,^{17,18} but we found that these approaches were not flexible enough to detect the underlying patterns present in our dataset, as they require a better-defined low-rank structure.

Our study had several other limitations, including decreased power due to limited sample size. Since data on $PM_{2.5}$ constituents were only available once every 3 or 6 days, our final dataset included only 978 observation days, despite spanning 9 years. Since our models had multiple covariates and nonlinear terms, a sample size of 978 may not have allowed for sufficient power to detect all associations that were present. Statistical power in the sensitivity analysis was further diminished, and the results of this analysis should be interpreted with caution. Furthermore, the noncontinuous sampling scheme did not allow use of distributed lag models to more robustly estimate lag-specific associations.

Our results are likely subject to exposure measurement error, as we did not have speciated $PM_{2.5}$ data available from all three monitors for each day included in analyses. We expect that $PM_{2.5}$ composition varies by geographic location, and we aimed to capture the city-wide values by taking the average of values measured at three separate locations. However, there were missing data in the AQS dataset for the Bronx monitor 2011–2014, and during most of 2007 for the Manhattan monitor. Our computed city-wide averages on days with missing data may not be comparable to computed city-wide averages on days with full data.

Conclusions

Applying PCP to speciated $PM_{2.5}$ data from the EPA's AQS database, and leveraging health outcome data from SPARCS, we found increased rates of hospital admissions for MI with increased same-day traffic, crustal dust, and chromium $PM_{2.5}$, as well as with lag 1 and 2 regional $PM_{2.5}$ in NYC from 2007 to 2015. To our knowledge, this is the first instance of applying PCP as a dimensionality-reduction method for speciated $PM_{2.5}$ data in an environmental epidemiology study. This study demonstrates the potential utility of PCP as a novel method for pattern recognition in environmental mixtures research.

Conflicts of interest statement

The authors declare that they have no conflicts of interest with regard to the content of this report.

References

- Hoek G, Krishnan RM, Beelen R, et al. Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environ Health*. 2013;12:43.
- Liu C, Chen R, Sera F, et al. Ambient particulate air pollution and daily mortality in 652 cities. *N Engl J Med*. 2019;381:705–715.
- Farhadi Z, Abulghasem Gorgi H, Shabaninejad H, Aghajani Delavar M, Torani S. Association between $PM_{2.5}$ and risk of hospitalization for myocardial infarction: a systematic review and a meta-analysis. *BMC Public Health*. 2020;20:314.

4. Brook RD, Rajagopalan S, Pope CA, et al. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. *Circulation*. 2010;121:2331–2378.
5. Newby DE, Mannucci PM, Tell GS, et al. Expert position paper on air pollution and cardiovascular disease. *Eur Heart J*. 2015;36:83–93b.
6. Weichenthal S, Lavigne E, Evans G, Pollitt K, Burnett RT. Ambient PM_{2.5} and risk of emergency room visits for myocardial infarction: impact of regional PM_{2.5} oxidative potential: a case-crossover study. *Environ Health*. 2016;15:46.
7. Dai L, Zanobetti A, Koutrakis P, Schwartz JD. Associations of fine particulate matter species with mortality in the United States: a multicity time-series analysis. *Environ Health Perspect*. 2014;122:837–842.
8. Davoodabadi Z, Soleimani A, Pourmoghaddas A, et al. Correlation between air pollution and hospitalization due to myocardial infarction. *ARYA Atheroscler*. 2019;15:161–167.
9. Ito K, Xue N, Thurston G. Spatial variation of PM_{2.5} chemical species and source-apportioned mass concentrations in New York City. *Atmos Environ*. 2004;38:5269–5282.
10. Lall R, Ito K, Thurston GD. Distributed lag analyses of daily hospital admissions and source-apportioned fine particle air pollution. *Environ Health Perspect*. 2011;119:455–460.
11. Kioumourtzoglou MA, Coull BA, Dominici F, Koutrakis P, Schwartz J, Suh H. The impact of source contribution uncertainty on the effects of source-specific PM_{2.5} on hospital admissions: a case study in Boston, MA. *J Expo Sci Environ Epidemiol*. 2014;24:365–371.
12. Rich DQ, Zhang W, Lin S, et al. Triggering of cardiovascular hospital admissions by source specific fine particle concentrations in urban centers of New York State. *Environ Int*. 2019;126:387–394.
13. Statewide Planning and Research Cooperative System (SPARCS). Available at: <https://www.health.ny.gov/statistics/sparcs/>. Accessed 27 November 2020.
14. US Environmental Protection Agency. Air Quality System Data Mart, Daily Summary Data: Particulates. Available at: <https://www.epa.gov/airdata>. Accessed 18 November 2020.
15. Cosgrove BA, Lohmann D, Mitchell KE, et al. Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDA) project. *J Geophys Res Atmospheres*. 2003;108D22:2002JD003118.
16. Candès EJ, Li X, Ma Y, Wright J. Robust principal component analysis? *J ACM*. 2011;58:1–37.
17. Gibson EA, Zhang J, Yan J, et al. Principal component pursuit for pattern identification in environmental mixtures. *Environ Health Perspect*. 2022;130:117008.
18. Zhang J, Yan J, Wright J. Square root principal component pursuit: tuning-free noisy robust matrix recovery. *Advances in Neural Information Processing Systems*. 2021;34:29464–29475.
19. R Core Team. R: A language and environment for statistical computing. Published online 2021. Available at: <https://www.R-project.org/>. Accessed 9 May 2022.
20. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. *BMC Bioinf*. 2010;11:367.
21. Wood SN. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC; 2006. Available at: <https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1910547>. Accessed May 10, 2022.
22. Masiol M, Hopke PK, Felton HD, et al. Source apportionment of PM_{2.5} chemically speciated mass and particle number concentrations in New York City. *Atmos Environ*. 2017;148:215–229.
23. Squizzato S, Masiol M, Rich DQ, Hopke PK. A long-term source apportionment of PM_{2.5} in New York State during 2005–2016. *Atmos Environ*. 2018;192:35–47.
24. Thurston GD, Ito K, Lall R. A source apportionment of U.S. fine particulate matter air pollution. *Atmos Environ*. 2011;45:3924–3936.
25. Zhou C, Zhou H, Holsen TM, Hopke PK, Edgerton ES, Schwab JJ. Ambient ammonia concentrations across New York State. *J Geophys Res Atmospheres*. 2019;124:8287–8302.
26. Thorpe A, Harrison RM. Sources and properties of non-exhaust particulate matter from road traffic: a review. *Sci Total Environ*. 2008;400:270–282.
27. Squizzato S, Masiol M, Rich DQ, Hopke PK. PM_{2.5} and gaseous pollutants in New York State during 2005–2016: spatial variability, temporal trends, and economic influences. *Atmos Environ*. 2018;183:209–224.
28. Dutkiewicz VA, Qureshi S, Khan AR, et al. Sources of fine particulate sulfate in New York. *Atmos Environ*. 2004;38:3179–3189.
29. Hopke PK, Zhou L, Poirot RL. Reconciling trajectory ensemble receptor model results with emissions. *Environ Sci Technol*. 2005;39:7980–7983.
30. Thurston GD, Burnett RT, Turner MC, et al. Ischemic heart disease mortality and long-term exposure to source-related components of U.S. fine particle air pollution. *Environ Health Perspect*. 2016;124:785–794.
31. Mills NL, Törnqvist H, Gonzalez MC, et al. Ischemic and thrombotic effects of dilute diesel-exhaust inhalation in men with coronary heart disease. *N Engl J Med*. 2007;357:1075–1082.
32. Mills NL, Törnqvist H, Robinson SD, et al. Diesel exhaust inhalation causes vascular dysfunction and impaired endogenous fibrinolysis. *Circulation*. 2005;112:3930–3936.
33. Ostro B, Lipssett M, Reynolds P, et al. Long-term exposure to constituents of fine particulate air pollution and mortality: results from the California Teachers Study. *Environ Health Perspect*. 2010;118:363–369.
34. Badaloni C, Cesaroni G, Cerza F, Davoli M, Brunekreef B, Forastiere F. Effects of long-term exposure to particulate matter and metal components on mortality in the Rome longitudinal study. *Environ Int*. 2017;109:146–154.
35. Franklin M, Koutrakis P, Schwartz J. The role of particle composition on the association between PM_{2.5} and mortality. *Epidemiology*. 2008;19:680–689.
36. Becker S, Dailey LA, Soukup JM, Grambow SC, Devlin RB, Huang YCT. Seasonal variations in air pollution particle-induced inflammatory mediator release and oxidative stress. *Environ Health Perspect*. 2005;113:1032–1038.
37. Nigra AE, Ruiz-Hernandez A, Redon J, Navas-Acien A, Tellez-Plaza M. Environmental metals and cardiovascular disease in adults: a systematic review beyond lead and cadmium. *Curr Environ Health Rep*. 2016;3:416–433.
38. Huang W, Cao J, Tao Y, et al. Seasonal variation of chemical species associated with short-term mortality effects of PM_{2.5} in Xi'an, a central city in China. *Am J Epidemiol*. 2012;175:556–566.