Canonical Correlation Analysis for Environmental Epidemiology: Air Quality Measures as Predictors of Cause-Specific County-Level Mortality

Erin Teeple
Data Science Program
Worcester Polytechnic Institute
Worcester, MA
eteeple@wpi.edu

Caitlin Kuhlman
Computer Science
Worcester Polytechnic Institute
Worcester, MA
cakuhlman@wpi.edu

Brandon Werner
Data Science Program
Worcester Polytechnic Institute
Worcester, MA
bwerner@wpi.edu

Randy Paffenroth
Mathematical and Data Sciences
Worcester Polytechnic Institute
Worcester, MA
rcpaffenroth@wpi.edu

Elke Rundensteiner
Data and Computer Sciences
Worcester Polytechnic Institute
Worcester, MA
rundenst@wpi.edu

***Abstract*—Many health outcomes result from multifactorial causal processes. These may be modelled using parametric methods which aim to evaluate the contributions of necessary and sufficient causes, for example among case and control subject populations. Where outcomes are not the deterministic results of singular events or exposures, however, we face the challenge of estimating parameters representing attributable fractions of risk for the outcome of interest, as well as making predictions from potentially incomplete and/or interrelated predictor exposure profiles. Adding to these challenges, parametric models may inherently lack the flexibility necessary to capture the true phenomena of interest. This project explores the application and interpretability of regression versus canonical correlation analysis (CCA) for studying cause-specific mortality risk resulting from air pollution exposure in the United States. We first create a novel exposure-outcome data set by integrating United States Environmental Protection Agency (EPA) annual summary county-level air quality measurements for the period 1980-2014 with age-adjusted gender- and cause-specific mortality rates from the same time period published by the Institute for Health Metrics and Evaluation. We then compare the performance of regression versus CCA models for predicting cause-specific county-level mortality from air quality measures.**

Keywords—air quality, environmental protection agency, EPA, exposure assessment, environmental epidemiology

I. INTRODUCTION

A. Background and Motivation

The United States Clean Air Act is a federal law that was first passed in 1970 and amended in 1977 and 1990 [1]. The Clean Air Act requires the United States Environmental Protection Agency (EPA) to set National Ambient Air Quality Standards (NAAQS) for six air pollutants termed "criteria air pollutants": ground-level ozone, particulate matter, carbon monoxide, lead, sulfur dioxide, and nitrogen dioxide[1]. Air pollution exposure has been previously linked with increased risk of adverse health events, including cardiac events and strokes [2-3]. In addition, negative health impacts have been found to be associated with air pollution exposure even at levels below current United States federal regulatory limits, for example in one study of a Medicare beneficiary population, increased all-cause mortality was found to be associated with higher levels of small-diameter particulate and ozone air pollution exposure that were within federal exposure limits [2]. Since air pollution exposure may impact morbidity and mortality risk across multiple organ systems, it becomes challenging to evaluate and quantify the effects of air pollution exposure on population health. Nonetheless, such dose-response and predictive models are necessary for evaluating and informing policies that regulate and update air pollution exposure limits.

One approach for modelling cause and effect relationships is the use of multiple linear regression, which estimates a response quantity from a set of predictor variables. For the situation where environmental air pollution exposure may impact multiple organ systems, however, the use of a multidimensional response vector has the potential to enable us to better capture the effects of interest, when these include altered rates of multiple potential adverse outcomes. In this work, we examine relationships among interrelated air pollution exposure measures and cause-specific mortality rates first as single rates, using multiple linear regression, and then by using Canonical Correlation Analysis (CCA), which finds combinations of predictor and response vector elements which are maximally correlated with each other, thereby permitting examination of intercorrelations among outcome variables with common predictors in the air quality exposure measures [4,5].

B. Related Work

Canonical correlation analysis (CCA) was described by Hotelling in 1936 [5]. CCA can be used when there exist multi-dimensional vectors $X = (x_1, x_2, \ldots x_n)$ and $Y = (y_1, y_2, \ldots y_n)$ in

which there exist correlations among variables [4,5]. This method seeks to find linear combinations of X and Y with maximal correlations with each other. In effect, these linear combinations may capture possible relationships between multidimensional X and Y domains [4,5].

Knapp has previously shown that significance tests for a number of popular statistical procedures (simple correlation, independent sample t test, multiple regression analysis, 1-way ANOVA, factorial ANOVA, analysis of covariance, correlated samples t test, discriminant analysis, and chi-squared test of independence) may be regarded as "special cases of the test of the null hypothesis in canonical correlation analysis for 2 sets of variables" [6]. Other recent works have further extended the core concepts of CCA to develop a representation-constrained CCA variation which is less sensitive to anomalous data [8] and to develop deep CCA, which is a nonlinear variant for application to detecting audiovisual synchrony [9].

C. Aims

This project applies canonical correlation analysis to characterize differential health effects of air quality using United States Environmental Protection Agency air quality monitor measurements obtained between 1980-2014. We compare the performance of CCA with traditional regression for characterizing relationships among our data fields and examine the statistical interpretations of our findings. Specific questions pursued in this exploratory analysis include how we may quantify relationships between air quality measures and population-level mortality and whether correlation methods may be used to better capture multidimensional health impacts.

II. Data Sets and Methods

A. Data Integration

AirData is a website maintained by the EPA that provides public access to air quality measurements collected at more than 4000 outdoor monitors across the United States, Puerto Rico, and the United States Virgin Islands. AirData has available for download annual and daily summary data tables containing measurements of criteria gases, particulates, meteorological conditions (wind, temperature, pressure, barometric pressure, and RH/dewpoint), toxics, ozone precursors, and lead measurements.

United States county-level age-standardized respiratory mortality rates for the years 1980-2014 are available through the Institute for Health Metrics and Evaluation [5]. The Institute for Health Metrics and Evaluation produced estimates for United States county-level mortality rates for 21 causes of death including chronic respiratory diseases for the period 1980-2014 [5]. This aggregated data set is available through the Global Health Data Exchange. Age-standardized mortality rates for male, female, and combined genders are reported as the number of deaths per 100,000 people in the population. These estimates were generated using death records from the National Center for Health Statistics (NCHS); population counts from the U.S. Census Bureau, NCHS, and the Human Mortality Database; and the cause list from the Global Burden of Disease Study (GBD). Web addresses for the public data sources used for this analysis are provided here:

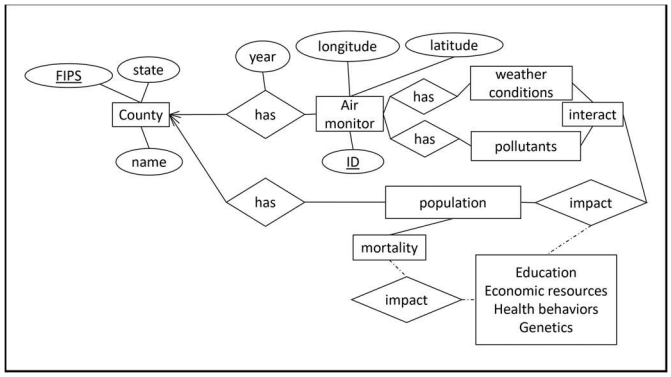

Fig. 1. Entity-relationship diagram guiding data integration.

(1) United States County Center Latitude and Longitude, FIPS codes; File: 2017_Gaz_counties_national.csv (attributes: FIPS, Latitude, Longitude); Source: https://www.census.gov/geo/maps-data/data/gazetteer2017.html

(2) Environmental Protection Agency AQI Annual Summary Files; Files: annual_aqi_by_county_[YEAR].csv; Source: https://aqs.epa.gov/aqsweb/airdata/download_files.html

(3) United States Combined and Gender-Specific Age-Adjusted Mortality Rates by United States County 1980-2014; Files: IHME_USA_COUNTY_MORTALITY_RATES_1980_2014_[STATE].csv; http://ghdx.healthdata.org/record/united-states-mortality-rates-county-1980-2014

B. Problem Definition and Methodology

Date pre-processing and table joins were implemented in Python, version 3.6, yielding a single .csv file containing 31,019 instances and 24 attributes. For this analysis, the 16-dimensional vector X includes year, county center latitude and longitude, and EPA air quality measures. The 8-dimensional vector Y is comprised of male and female age-adjusted mortality rates for four causes: Respiratory disorders (RESP), Cardiovascular diseases (CVD), Traumatic injuries (INJ), and Neurological disorders. These causes were selected due to the strength of the reported associations between adverse cardiovascular and respiratory effects and air pollution exposure, while deaths due to traumatic injuries and neurological disorders were selected as comparison groups. Pearson correlations are presented in Fig. 2.

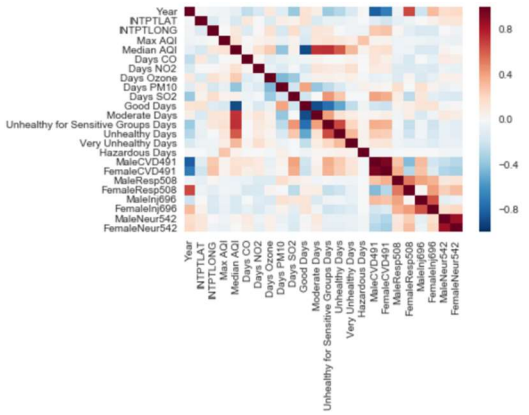

Fig. 2. Pearson correlations for air quality and mortality variables

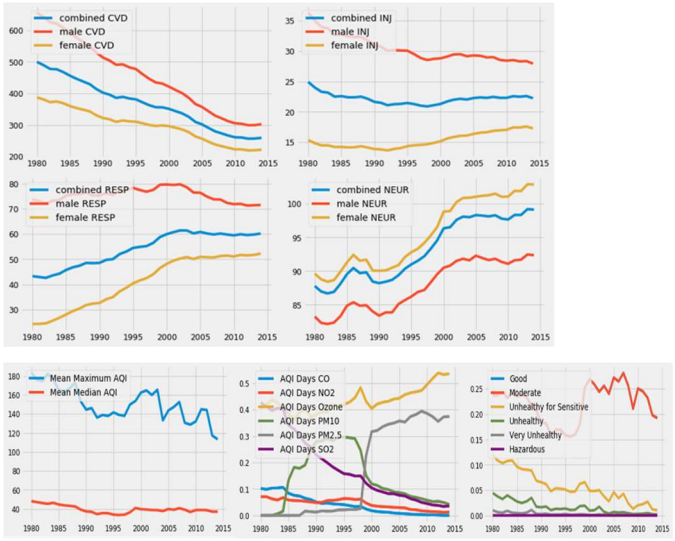

Fig. 3. United States mean mortality rates and air quality markers 1980-2014

Initial data exploration was performed to characterize temporal trends in the mortality and air quality variables. Different temporal trends for each of the mortality causes were observed, and air quality measures were also observed to vary during the study period (Fig. 3).

Due to the observed temporal variability, year was included in all of our linear models. We included latitude and longitude in our integrated data set, as well, given that we hypothesized that baseline mortality levels might vary by region. With respect to our air pollution exposure measurements, several aspects of these variables limited their use together in regression. Air Quality Index (AQI) is summary measurement of air quality, with scores from 0-500 corresponding to ratings: Good (0-50), Moderate (51-100), Unhealthy for Sensitive Groups (101-150), Unhealthy (151-200), Very Unhealthy (201-300), and Hazardous (301-500). Limitations of this scoring system include that the score itself does not identify what specific pollutants account for the overall score. While our data set does include the proportion of days on which a specific pollutant accounted for the maximal AQI score, this does not allow us to identify co-exposures to other air pollutants at graded levels. Additionally, since median AQI, maximum AQI, and proportion of days on which AQI was good or moderate are inherently interrelated, multiple linear regression models were trialed including each of these individually, with and without interaction terms between the air quality measures and the proportion of days on which a specific pollutant accounted for the AQI, in an effort to capture the interaction between the leading pollutants and the magnitude of exposure at the time and location of each observation point. Multiple linear regression and canonical correlation analysis was performed in R using the function cancor and the CCP package for significance testing. Visualizations of test and training data for the first two X and Y canonical components were generated in Python [7] and R [11].

III. Results

A. Multiple Linear Regression

Table 1 presents the Adjusted R^2 and Mean-Squared Error (MSE) for regression performed using location and year only

TABLE I. Multiple Linear Regression

Mortality Rate	*Air Quality Measures Included in Multiple Linear Regression without Interaction Terms*							
	No Air Quality Attributes		Proportion of Good/Moderate Days		Median AQI		Maximum AQI	
	AdjR²	MSE	AdjR²	MSE	AdjR²	MSE	AdjR²	MSE
MaleCVD	0.74	4110	0.70	4812	0.70	4770	0.70	4816
FemaleCVD	0.58	2010	0.52	2334	0.52	2311	0.51	2344
MaleRESP	0.01	302	0.06	288	0.06	288	0.06	288
FemaleRESP	0.44	104	0.44	104	0.44	104	0.44	104
MaleINJ	0.07	57	0.11	55	0.11	55	0.10	56
FemaleINJ	0.12	11	0.13	11	0.13	11	0.13	11
MaleNEUR	0.04	359	0.05	355	0.06	355	0.06	355
FemaleNEUR	0.06	528	0.06	524	0.06	524	0.07	522
Mortality Rate	*Air Quality Measures Included in Multiple Linear Regression with Interaction Terms*							
	No Air Quality Attributes		Proportion of Good/Moderate Days		Median AQI		Maximum AQI	
	AdjR²	MSE	AdjR²	MSE	AdjR²	MSE	AdjR²	MSE
MaleCVD	*as above*	*as above*	0.70	4788	0.70	4756	0.70	4796
FemaleCVD			0.52	2314	0.52	2308	0.52	2331
MaleRESP			0.06	287	0.06	287	0.06	287
FemaleRESP			0.44	104	0.44	104	0.44	104
MaleINJ			0.12	54	0.11	55	0.12	54
FemaleINJ			0.14	11	0.13	11	0.13	11
MaleNEUR			0.06	355	0.06	353	0.06	353
FemaleNEUR			0.06	522	0.07	521	0.07	520

'No Air Quality Attributes' includes year, latitude, and longitude as predictors; All other regressions include with air quality variables for the proportions of days when AQI was attributed to NO2, Ozone, SO2, and PM10, as well as interaction terms between these and the air quality measures in the lower table; Air quality regressions do not include latitude and longitude; AdjR²- Adjusted R-squared; MSE - mean-squared error; F-statistic p-values < 0.05 for all regressions.

to predict mortality and comparison models where year and air quality (without location) were used to predict mortality rates for cause and gender. Remarkably, air quality measures achieved similar performance to location (latitude and longitude) for predicting mortality rates by cause. We observe that the highest R^2 for all models is for the prediction of male cardiovascular mortality from location or air quality. From this data set, it cannot be inferred the air pollution is the cause of location-specific variations in male cardiovascular mortality, however the close performance of the models is notable. We did not find that the inclusion of interaction terms improved performance of our models, but given that our data set does not include all pollutant exposures and levels for each AQI score, the pollutant data fields may not have accurately captured the actual pollutant exposures so that we could examine these in our model.

B. Canonical Correlation Analysis

Using year, air quality measures, and latitude and longitude as eleents in the X vector, CCA was performed separately for two Y vectors: i) male and female cardiovascular and respiratory mortality and ii) male and female injury and neurological disorder mortality. Fig. 4 presents the cross-correlation matrices for these analyses.

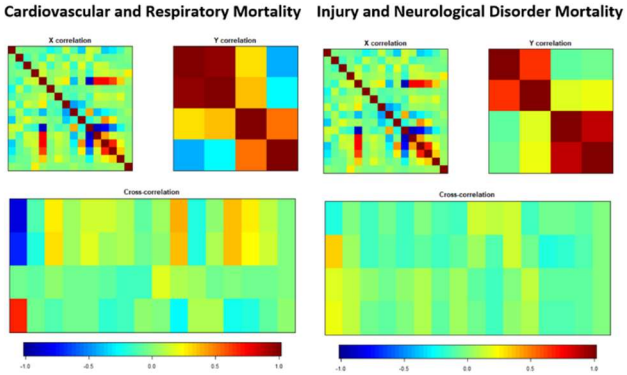

Fig. 4. X, Y, and Cross-Correlations by mortality rate groupings

TABLE II. CANONICAL COEFFICIENTS

Y – Cardiovascular and Respiratory Mortality

	[,1]	[,2]	[,3]	[,4]
year	-1.010730e-01	-0.0245142848	1.729825e-03	-0.0095281346
LAT	-1.075018e-02	0.0857233761	-1.071644e-01	-0.0403118786
LONG	1.227109e-02	-0.0262772289	5.275648e-03	-0.0109774355
MAX_AQI	-4.519864e-05	-0.0002733604	-4.452063e-04	0.0003820838
MED_AQI	2.507747e-03	0.0255786655	1.497807e-02	-0.0090275972
pdaysCO	-2.627648e-01	0.6454981986	-2.851815e+00	-3.5123108980
pdaysNO2	-2.869132e-01	-2.1297800426	-3.400280e+00	5.0520814321
pdaysOZONE	-1.230335e-01	-0.3354489435	-7.057615e-01	0.6741533260
pdaysPM10	-1.126992e-01	0.6603625074	1.414683e+00	0.2895303618
pdaysSO2	1.535546e-01	0.3476389516	7.509393e-01	-0.4958458132
pAQI_GOOD	-6.784995e+00	24.8342568989	-1.678234e+01	25.2620685344
pAQI_MOD	-6.926810e+00	22.0790348424	-1.682564e+01	22.3124875958
pAQI_UNHS	-6.946765e+00	19.9517059360	-1.860801e+01	23.8357471604
pAQI_UNH	-7.212052e+00	20.1567096423	-2.148714e+01	36.8715388547
pAQI_VUNH	-6.971732e+00	7.1958276785	-2.072798e+01	7.8819215079
pAQI_HAZ	-8.148781e+00	70.9003703821	-5.436010e+01	78.2208041642

Y – Injury and Neurological Disorder Mortality

	[,1]	[,2]	[,3]	[,4
year	-1.059462e-01	-1.058075e-02	3.245995e-02	-4.211299e-0
LAT	-3.085789e-02	-1.694688e-02	-1.010485e-01	-1.328765e-0
LONG	-2.756395e-03	-3.800654e-02	-3.923972e-04	1.800250e-0
MAX_AQI	-8.011970e-05	-5.902593e-04	8.437460e-04	2.094977e-0
MED_AQI	5.874951e-03	3.969477e-03	-6.297320e-03	3.296992e-0
pdaysCO	-8.518102e-01	-2.445447e+00	-3.627256e+00	-2.080596e+0
pdaysNO2	-6.953006e-01	-3.952872e+00	2.871634e+00	-6.526437e-0
pdaysOZONE	-1.263039e-01	-9.716171e-01	-4.684436e-01	-3.136360e+0
pdaysPM10	-1.218877e-02	-3.160100e-01	-3.310942e-02	-2.750968e+0
pdaysSO2	1.700409e-01	1.986288e-01	8.420553e-01	-1.807725e+0
pAQI_GOOD	-2.944719e+01	-2.531408e+01	-5.139097e+01	-1.476594e+0
pAQI_MOD	-3.045842e+01	-2.649999e+01	-5.397582e+01	-1.488410e+0
pAQI_UNHS	-3.057121e+01	-2.782551e+01	-4.891519e+01	-1.483631e+0
pAQI_UNH	-3.011634e+01	-3.094667e+01	-5.036181e+01	-1.649796e+0
pAQI_VUNH	-3.302387e+01	-4.023427e+01	-3.346611e+01	-1.566338e+0
pAQI_HAZ	-3.137791e+01	-2.800729e+01	-2.844817e+01	-2.696114e+0

In both mortality groupings, tests of the canonical dimensions for the CCA analyses lead us to reject the null hypothesis that the canonical correlations in the current row and all that follow are zero for all rows. We may also examine which variables most strongly influence the canonical dimensions by examining their weightings. These are presented in Table 2. A visualization of the fits of the relationships between the canonical X and Ycomponents in the CCA modelis shown for cardiovascular and respiratory mortality in Fig. 5 and for injuries and neurological disorders in Fig. 6. Note the component 1 linearity achieved for combined cardiovascular and respiratory mortality.

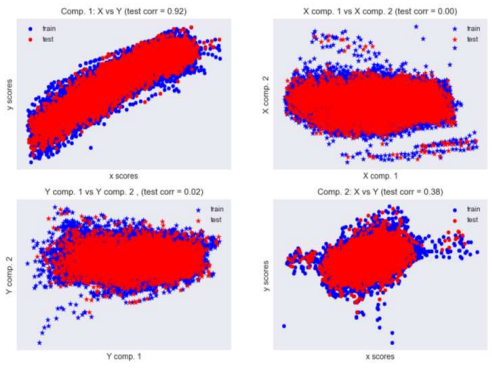

Fig. 5. Plots of test and training data for the first two X and Y canonical components where Y is composed of male and female cardiovascular mortality rates. Note the linearity achieved for component 1 X vs. Y.

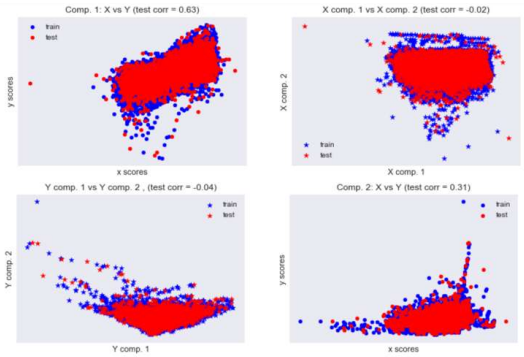

Fig. 6. Plots of test and training data for the first two X and Y canonical components where Y is composed of male and female injury and neurological disorder mortality.

IV. CONCLUSIONS

Intercorrelations among our air quality exposure measures limit our ability to use multiple linear regression to explore associations between air pollution exposure and mortality in this data set. Canonical correlation analysis, which does not require predictor independence, or a single target outcome achieves high first-component test correlation when used to model relationships among annual summary air quality measures and cardiovascular and respiratory mortality. These results suggest that CCA-based methods offer promise for environmental epidemiology applications.

Acknowledgment

We thank the WPI Data Science Program for feedback and support on this project.

REFERENCES

[1] https://www.epa.gov/laws-regulations/summary-clean-air-act; 1/2019.

[2] Q. Di, Y. Wang, A. Zanobetti, others. “Air pollution and mortality in the Medicare Population” *NEJM,* 2017, 376, (26).

[3] A. Shah, K. Lee, D. McAllister, others. “Short term exposure to air pollution and stroke: systematic review and meta-analysis” *BMJ, 2015, 24,350.*

[4] W. Härdle, L. Simar. "Canonical Correlation Analysis". *Applied Multivariate Statistical Analysis*. 2007. pp. 321–330. doi:10.1007/978-3-540-72244-1_14. ISBN 978-3-540-72243-4.

[5] H. Hotelling. "Relations Between Two Sets of Variates". *Biometrika, 1936,28 (3–4):321377. doi:10.1093/biomet/28.34.321. JSTOR 2333955.*

[6] T.R. Knapp. "Canonical correlation analysis: A general parametric significance-testing system". *Psych Bull, 1978,* 85, (2): 410–416.

[7] https://scikitlearn.org/stable/auto_examples/cross_decomposition/plot_compare_cross_decomposition.html#sphx-glr-auto-examples-cross-decomposition-plot-compare-cross-decomposition-py

[8] Representation-Constrained Canonical Correlation Analysis: A Hybridization of Canonical Correlation and Principal Component Analyses. *Journal of Applied Economic Sciences*, 2009, 4(1), 115–124.

[9] S. Sieranoja, M. Sahidullah, T. Kinnunen, others. “Audiovisual synchrony detection with optimized features” http://cs.joensuu.fi/pages/tkinnu/webpage/pdf/audiovisual_synchrony_2018.pdf.

[10] S-Y. Huang, M-H. Lee, K.Chuhsing. “Kernel Canonical Correlation Analysis and its Applications to Nonlinear Measures of Association and Test of Independence”. *Journal of Statistical Planning and Inference*, 2009, 139, (7):2162.

[11] I. Gonzalez, S. Dejean, P. Martin, A. Baccini. “CCA: An R package to extend canonical correlation analysis”. *J Stat Software* 2008, 23 (12).