Studying the Association of Environmental Measures Linked with Health Data: A Case Study Using the Linked National Health Interview Survey and Modeled Ambient PM2.5 Data

Rong Wei[1], Van Parsons, and Jennifer Parker

National Center for Health Statistics, Hyattsville, MD 20782

Abstract

Studies have shown that air quality is associated with population health. Health and air data are collected by two independent data systems, the National Health Interview Survey (NHIS) and the Air Quality System (AQS), respectively. To overcome the limited spatial-time coverage of AQS monitor data, an extensive model-predicted universe of spatial-time air measurements was created. These approaches were developed by the Environmental Protection Agency (EPA) and adopted for use in public health by CDC's Environmental Public Health Tracking Program, National Center for Environmental Health (NCEH). From this universe, air measurements for PM2.5 and ozone were linked at the census tract level to the NHIS sample over years 2001 to 2010 for those areas within the contiguous-US. As this linkage is complete, air/health association analyses on the NHIS can be performed using suitable design-based or model-based methods in contrast to analyzing a partial air/heath linkage with the original AQS air data. This study is somewhat exploratory with the needs of a typical NHIS data user in mind. The study attempts to determine some of the operating characteristics of the linked data and presents suggestions for design-based and model-based analyses. In particular, some basic spatial-time associations are explored. Some thoughts on next steps are also discussed at the end of the paper.

Keywords: air pollutant, health status, modeled PM2.5 measures, linked complex survey data, mixed effects model, model-based analysis

[1] *The findings and conclusions in this study are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.*

1. Introduction

Linking health data with other data sources enhances the depth of analytical studies on health related issues (Parker *et al*. 2009). Within the Centers for Disease Control and Prevention (CDC) a number of national surveys, administrative data, as well as environmental data have been linked to air quality data (Parker *et al.* 2008a; 2008b; Robert *et al*. 2014). Appropriate methods for statistical analysis of survey and linked data involve many aspects, for example, determining the completeness and accuracy of the matched units and determining suitable modified design structures. These issues tend to be data and project specific (Wei and Parsons, 2009; Judson *et al*. 2013; Parsons *et al*. 2013). In a previous study, Wei, *et al.* (2014) focused on design features based on an incomplete linkage of NHIS data with EPA air quality monitor data. Air quality monitors only sparsely cover US geography; depending upon the air quality unit only about 20-80% of NHIS sampled persons will have linked EPA information. Consequently, any direct NHIS population-based inference using the directly linked monitor data may be subject to large biases. An alternative approach is to use an air quality prediction model that covers US geography. Having such a model, (assumed to be reasonably accurate and capable of providing timely predictions), can allow complete linkage of air quality measurements to NHIS sampled geography and thus allow the complete NHIS design structures to be used for analyses.

Such a model has been developed under an EPA sponsored project, the Community Multi-scale Air Quality (CMAQ) modeling system, and this model has been adapted for linkage by CDC's Environmental Public Health Tracking Program with several CDC data systems. (Some discussion is provided in Section 2.1 below.) The final model used in this study provided estimates of daily air quality for PM2.5 (particulate matter with diameter of 2.5 micrometers or less, also called fine particles) and ozone for all Census tract areas in the contiguous US (Alaska and Hawaii excluded). While a linkage of this air quality data to the sampled NHIS areas results in complete geographical coverage, issues in handling the linked daily air measurements with an NHIS interview at a single point time must be resolved. Furthermore, while the data are modeled at the tract level, for an analysis it may be advantageous to aggregate at a coarser level, e.g., the county level. In this paper some of the basic analytic issues for studying associations between reported health and air quality are explored with only preliminary results displayed. Substantive findings will not be discussed in this paper as the objective of the paper is methodological. A more comprehensive report will be a future endeavor.

2. Data and their linkages

2.1 Description of Air Quality Data

For the PM2.5 modeling, ambient PM2.5 data during years 2001 to 2010 were collected from about 1100 national monitors at specified time intervals through the US EPA's Air

quality System (AQS). The CMAQ modeling project modeled the raw air quality estimates by first "fusing" monitor data into gridded outputs and then fitting a hierarchical Bayesian (HB) model. Using these modeled PM2.5 values as a starting point, the NCEH further modified the CMAQ modeled data to cover all geographical census tract levels in the contiguous US (over 70,000 areas). Similar methods were used to model ozone measurements. (See Vaidyanathan *et al*. 2013 for details and additional references on AQS and CMAQ modeling procedures.) Compared with the original monitor data, the modeled data have the advantages of nation-wide coverage at a fine-level geographical unit, "no recording time gaps" during a year (daily estimates for 365 days), and having a census tract ID available for data linkage. As a resource, it should be kept in mind that the modeled data are not deterministic, but subject to variance and bias. This preliminary study has not studied these issues.

2.2 Description of the NHIS

Contrasted to the EPA monitored data, the NHIS data are based on a complex survey design whose estimates represent the civilian non-institutionalized population of the United States. The NHIS survey data include survey weights and clustering factors that are recommended for design-based data analysis.

2.3 Linkage of the NHIS and Air Quality Data

NHIS data covering survey years 2001 to 2010 were linked to modeled PM2.5 data at the 2010-defined census tract level. Table 1 shows the scope of the linked data for air quality 2005 data. Here, the NHIS has 98,649 individual records, covering 8470 census tracts within 844 US counties; all units are linked to modeled PM2.5 data. If only the original monitored data are considered, then 19% of the data have no linkage, thus making standard design-based analyses using the NHIS problematic.

Table 1. Overview scope of linked NHIS and modeled PM2.5 data, 2005

<table>
<tr><td colspan="5">Year2005</td><td colspan="2">Time Detail</td></tr>
<tr><td>Modeled PM2.5 data</td><td>#Counties</td><td>#Census Tracts</td><td>#NHIS individuals</td><td>PM2.5 monitor data</td><td>Modeled PM2.5</td><td>NHIS</td></tr>
<tr><td rowspan="2">Areas have NHIS samples</td><td rowspan="2">844</td><td rowspan="2">8470</td><td rowspan="2">98,649</td><td>80,109 individuals from 626 counties have monitor data</td><td rowspan="3">Daily measures, seasonal and annual averages</td><td rowspan="2">One time Interview, IDs for day, season, and year</td></tr>
<tr><td>18,540 individuals have no monitor data</td></tr>
<tr><td>Areas have no NHIS samples</td><td>2299</td><td>63,813</td><td></td><td></td><td></td></tr>
</table>

The modeled PM2.5 measurement time series has been averaged into four quarter values as well as annual values for the linkage. These measurement times are consistent with the NHIS interview process by annual quarter. The NHIS individual records are recorded at a one-time interview at a day within a quarter and year.

Since the original two independent data collection designs have different geographical and population coverages, we first explored the distributions of modeled PM2.5 between census tracts with and without NHIS samples. Figure 1 shows that the distributions of PM2.5 across years and seasons are very similar between areas with and without NHIS samples. Since the ratio of the number NHIS tracts to the number of non-NHIS tract coverage is roughly 1 to 8, and NHIS tracts are in the sample proportional to population size, this figure is similar to plotting PM2.5 in high population-level to low population-level tracts. This pattern is somewhat consistent with the pattern of PM2.5 for metro and non-metro regions displayed later in Figure 4.

It should be noted that the annual modeled PM2.5 estimates during 2006-2010 used for this study were lower than the National Ambient Air Quality primary and secondary standards, 12 and 15 μg/m3, respectively.

3. Usage of NHIS design factors with linked data

3.1 Modeling Structures

Treating the modeled PM2.5 as an accurate deterministic measure, design-based analyses are possible for population-based inference. For design-based methods to be statistically stable, usually coarse, well-sampled subdomains are required. Whenever domains of interest become sparely sampled, in particular for smaller geographical domains, or for analyses that are focused on relationships among variables, model-based analysis may be an effective alternative due to their flexibility in defining structure.

To use model-based analyses with complex survey data, it recommended that the survey structures of weighting and clustering be incorporated into the modeling. For the NHIS, the variables of race/ethnicity are used to define NHIS substrata and define differential sampling rates, and these variables along with gender and age are used to adjust sampling weights. It is suggested that these three types of NHIS variables be used as fixed effects covariates in any modeling. NHIS clustering is hierarchical with multiple levels. The first random cluster level is the primary sampling unit (PSU). This cluster can be treated as a random effect in analyses. More complicated random effects models involving finer geographically defined clusters along with household clusters can be used, but the many imbalances require more effort in setting up effective models. Only the PSU-level random effect was considered in the preliminary analyses.

Models may be constructed as weighted and unweighted at the individual level. The NHIS survey weights can be scaled to reflect an effective sample size, (discussed in Section 4.4

of Korn and Graubard (1999)); this modified weight can be applied as a weighting factor in modeled analyses. This weight modification technique seems to help in accounting for the survey design. In Wei and Parsons (2009), using the scaled individual weight, defined by $wtsca_i \equiv n_{total} (w_i / \sum_j w_j) / (CV^2(w) + 1)$, where the w's are NHIS survey weights for individuals, and n_{total} is the unweighted total on the data domain targeted for analysis, was shown to provide some degree of agreement when comparing design- and model-based inference on larger domains.

Using the structures just discussed, at least as a starting point, standard modeling packages can be used for complex survey analysis.

3.2 Modeling Examples for Health Status with PM2.5 and Ozone as Covariates

For preliminary analyses the following variables were considered:

Response:
healthStatus:
healthy(0) , reported health status excellent, very good, and good;
unhealthy (1), reported health status fair or poor.

Covariates:
race_ethnic: Hispanic, non-Hispanic white, non-Hispanic black and others;
gender: male and female;
age: continuous age 0-85 and centered at 45;
Ozone and PM2.5: continuous, annual tract averages, centralized.

Weight:
wtfa: NHIS survey weight;
wtsca: wtfa scaled to effective sample size.

Design Information:
strata and PSU

SAS analyses and codes:

1) Design-based with SAS®SurveyLogist:

Proc Surveylogistic;

Class race_ethnic gender /param=glm;

Stratum strata;

Cluster PSU;

Model healthStatus = race_ethnic gender age Ozone PM2.5;

Weight wtfa;

2) Simple model using SAS®GLIMMIX, but no weight or random effects:

Proc Glmmix;

Class race_ethnic gender;

Model healthStatus = race_ethnic gender age Ozone PM2.5/**dist**=binary **link**=logit solution;

3) Mixed effects model using SAS®GLIMMIX with weight and random effects:

Proc Glimmix;

Class race_ethnic gender**;**

Model healthStatus = race_ethnic gender age Ozone PM2.5/**dist**=binary **link**=logit solution;

Random int/subject=PSU;

Weight wtsca;

Select outputs from these runs are displayed in Figure 2.

Note, the mixed effects model presented above makes the most use of the survey design features of clustering, differential weighting by race/ethnicity and poststratification. Interactions among covariates were also explored, but in general, no significant results were seen, and so all final models at this preliminary stage did not include interaction terms.

As can be seen in Figure 2, the design-based and simple model-based methods track each other fairly well. The mixed effects model's pattern appears somewhat different than other two approaches before year 2007, but more similar after 2007. However, only the year 2010 PM2.5 effects are statistically significant.

4. Exploring associations by metro-status and seasonality

The results of section 3 are those from the complete data in which geographical areas are not distinguished, and the air pollutants are not time specific, but annual tract-level averages. For exploration of these spatial-time variables, displays of PM2.5 data distributions, by location and time can be seen in Figures 3, 4 and 5. These figures suggest that geography and PM2.5 seasonality, along with associated population concentrations by metropolitan, micropolitan, and nonmetropolitan status may be informative in the study of associations of health and PM2.5. (The urbanization definitions were based on OMB year 2000 standards applied to years 2003 to 2009. A more refined definition developed by NCHS will be used in future work.) Figure 3 shows that the PM2.5 measures vary across US geographic regions. Figure 4 shows that PM2.5 measures were higher in

metropolitan and nonmetropolitan areas than in micropolitan areas over years 2003 to 2009. Figure 5 shows that over four seasons, PM2.5 tends to be higher in the season of July to September compared to other three seasons.

Several exploratory association analyses were applied using metro status and seasonality factors. As the data were partitioned into subsets for these analyses, some design structure needed for design-based analyses was lost. Mixed effects models were used to overcome the deficiencies. The results are shown in Figures 6 and 7. It can be seen in Figure 6 that the PM2.5 effects on health status are significant in non-metropolitan areas for all years of 2003 to 2009, but not in the other two types of areas. In Figure 7, it can be seen that the PM2.5 measured in the summer season (July-September) have significant effects on health status in five (2003, 2004, 2006, 2007 and 2010) years out of 10 years.

5. Further study

Analysis with linked NHIS survey data and modeled AQS data involves many issues which affect study results. Due to space constraints, this paper only considered a few of these many issues. Future work should give attention to using finer structured random effects models that include county and tracts components of variance along with PM2.5 spatial-time variables modeled within the entire 2001-2010 data system. Further study will also explore some underlying assumptions about the modeled PM2.5. For an example, do direct NHIS inflation estimates, applied to the modeled PM2.5 reflect the US population on smaller domains? Currently, the modeled PM2.5 data are treated as deterministic when the PM2.5 data are actually modeled Bayesian posterior means with posterior variances. Approaches for incorporating this variance need to be studied.

Acknowledgement

Sincere thanks to Ambarish Vaidyanathan at National Center for Environmental Health, CDC for providing the modeled Air Quality Data used in this study.

References

Korn, E, Graubard B (1999). Analysis of Health Surveys: Wiley.

Judson DH, Parker JD, Larsen MD. (2013). Adjusting sample weights for linkage-eligibility using SUDAAN. National Center for Health Statistics, Hyattsville Maryland. Available at the following address: http://www.cdc.gov/nchs/data/datalinkage/adjusting_sample_weights_for_linkage_eligibility_using_sudaan.pdf

Parker JD, Kravets N, Woodruff TJ. (2008a). Linkage of the National Health Interview Survey to air quality data. National Center for Health Statistics. Vital Health Stat(145):1–24.

Parker DJ, Woodruff TJ, Akinbami LJ, Kravets N. (2008b). Linkage of the US National Health Interview Survey to air monitoring. Environmental Research 106 384–392.

Parker DJ, Woodruff TJ, Akinbami LJ. (2009). Air Pollution and Childhood Respiratory Allergies in the United States. Environmental Health Perspectives, Vol 117 (1) 139-147.

Parsons V, Wei R, Parker JD. (2013). Evaluation of Model-based Methods in Analyzing Complex Survey Data: A Simulation Study using Multistage Complex Sampling on a Finite Population. ASA Proceedings of the Joint Statistical Meetings, pp 3446-3456.

Robert JD, Voss JD, Knight B. (2014). The Association of Ambient Air Pollution and Physical Inactivity in the United States PLoS ONE, 9(3): e90143. http://doi.org/10.1371/journal.pone.0090143

Vaidyanathan A, Dimmick WF, Kegler SR, Qualters JR. (2013). Statistical air quality Predictions for Public Health Surveillance: Evaluation and Generation of County Level Metrics of PM2.5 for the Environmental Public Health Tracking Network. International Journal of Health Geographics 12:12.

Wei R, and Parsons V. (2009). Model-based Methods in Analyzing Complex Survey Data: A Case Study with National Health Interview Survey data. ASA Proceedings of the Joint Statistical Meetings, pp 2558-2567.

Wei R, Parsons V, Parker J and He Y, (2014). Data Analysis Using NHIS-EPA Linked Files: Issues with Using Incomplete Linkage. ASA Proceedings of the Joint Statistical Meetings, pp 3203-3213.

Figure 1. Estimated annual average PM2.5 overall and during July-Sept and estimated annual maximum for areas with and without NHIS samples, by year.
NCEH HB air quality data 2001-2010.

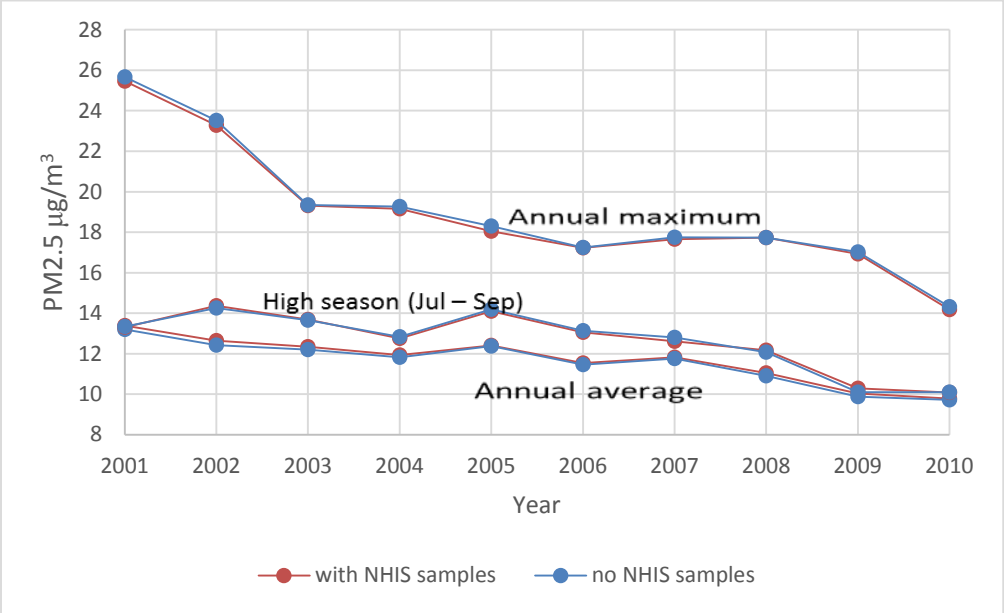


Figure 2. Association between PM2.5 and reported health status (β estimate) using one design-based and two model-based approaches, by year.
2001-2010 NHIS linked to NCEH HB air quality data.

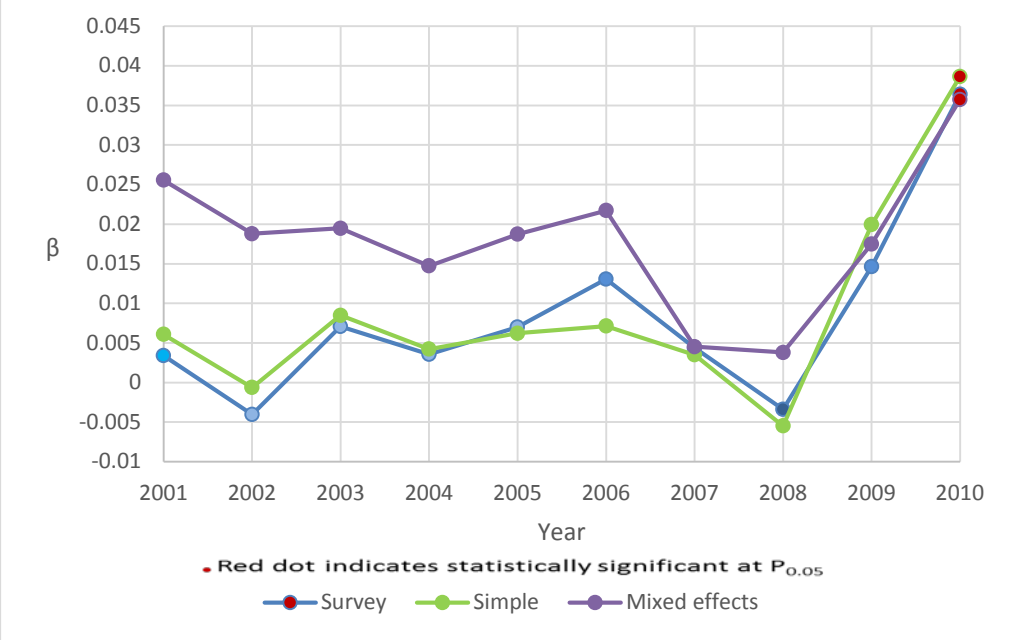


Figure 3. Annual modeled PM2.5 distribution over contiguous US: 2001 and 2010

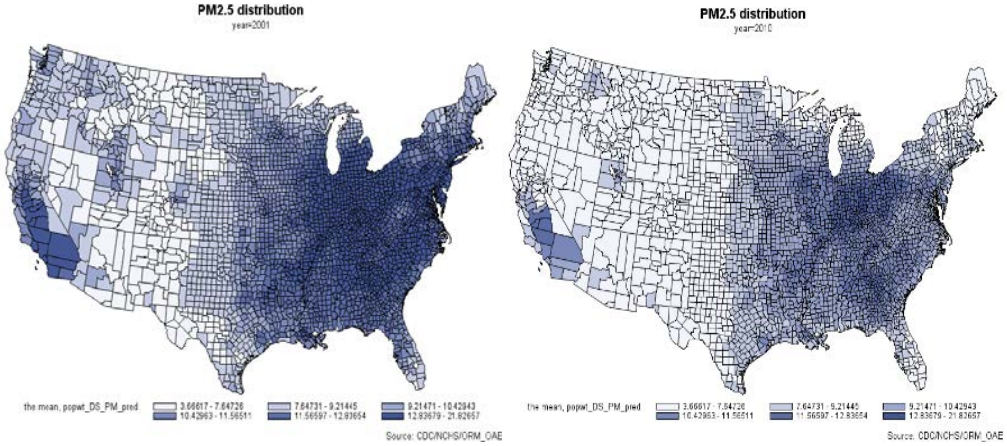
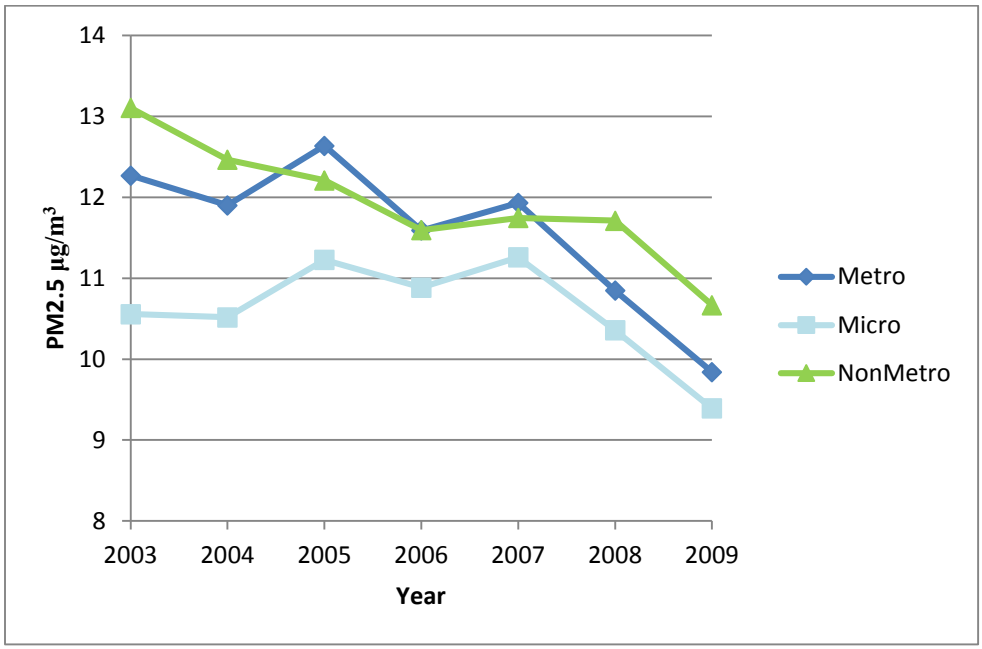


Figure 4. Annual modeled PM2.5 in metropolitan, micropolitan and nonmetropolitan areas over the contiguous US, 2003 – 2009.



Urbanization Population ranges in 1000's
Metro: 100+, Micro: 50 to 100, Non Metro: less than 50

Figure 5. Seasonal PM2.5 modeled measures in the contiguous US, 2001 - 2010

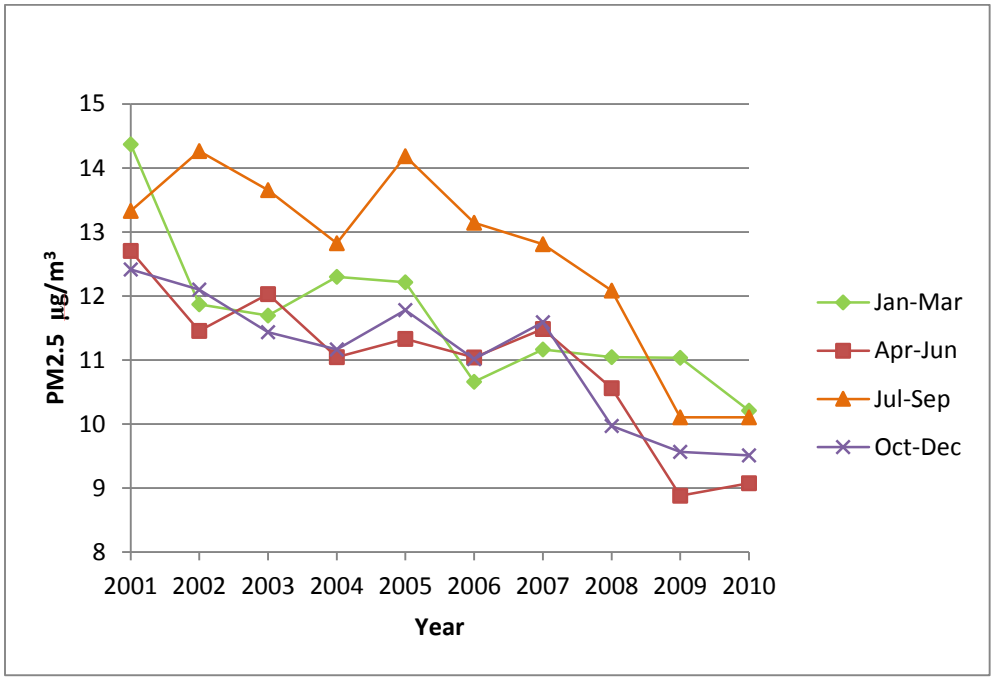
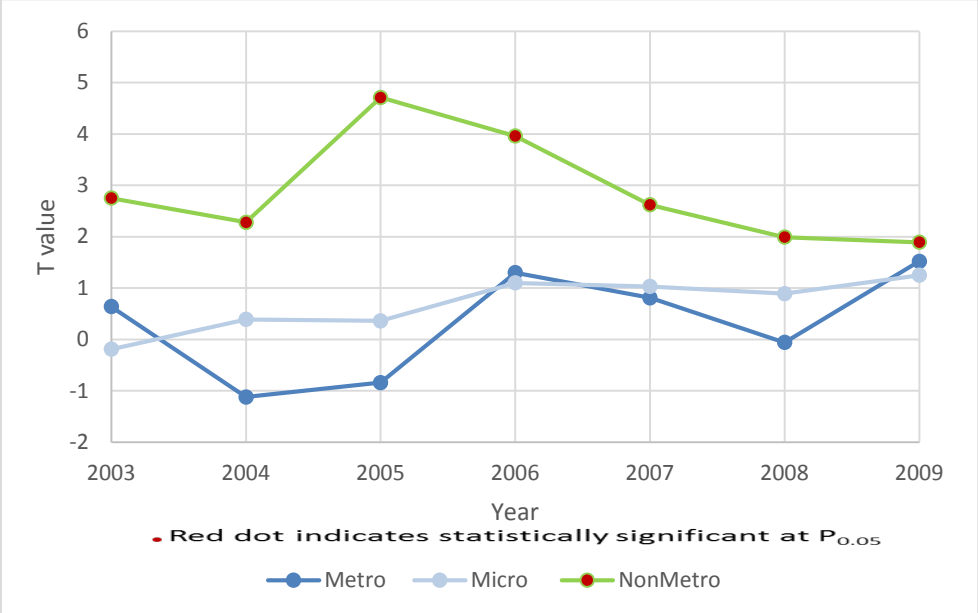


Figure 6. Association between PM2.5 and reported health status (t-statistic) using mixed effects model, by year and level of urbanization. 2003-2009 NHIS linked to NCEH HB air quality data.



Urbanization Population ranges in 1000's
Metro: 100+, Micro: 50 to 100, Non Metro: less than 50

Figure 7. Association between PM2.5 and reported health status (t-statistic) using mixed effects model, by year during the high PM2.5 season (July – September). 2001-2010 NHIS linked to NCEH HB air quality data.