On Ecological Fallacy and Assessment Errors Stemming from Misguided Variable Selection: Investigating the Effect of Data Aggregation on the Outcome of Epidemiological Study

Boris A. Portnov Ph.D., DSc[1], Jonathan Dubnov MD, MPH[2,4] and Micha Barchana MD, MPH[3, 4]

Institution: University of Haifa, Israel

[1]Department of Natural Resources & Environmental Management, University of Haifa, Israel

[2]Haifa District Health Office, Ministry of Health, Israel

[3] Public Health Services, Ministry of Health, Israel

[4] School of Public Health, University of Haifa, Israel

Address for correspondence: Dr. Jonathan Dubnov, Haifa District Health Office, Ministry of Health, 15a Palyam Ave., Haifa, 31999, Israel. E-mail: jonathan.dubnov@lbhaifa.health.gov.il

Table of Contents

Abstract

In social and environmental sciences, ecological fallacy is an incorrect assumption about an individual based on aggregate data for a group. In the present study, the validity of this assumption was tested using both individual estimates of exposure to air pollution and aggregate data for 1,492 schoolchildren living in the in vicinity of a major coal-fired power station in the Hadera region of Israel. In 1996 and 1999, the children underwent subsequent pulmonary function tests (PFT), and their parents completed a detailed questionnaire on their health status and housing conditions. The association between children's PFT results and their exposure to air pollution was investigated in two phases. During the first phase, PFT averages were compared with average levels of air pollution detected in small census areas in which the children reside. During the second phase, individual pollution estimates were compared with individual PFT results, and pattern detection techniques (Getis-Ord statistic) were used to investigate the spatial data structure. While different levels of areal data aggregation changed the results only marginally, the choice of indices measuring the children's PFT performance had a significant influence on the outcome of the analysis. As argued, a difference between individual and group correlations (i.e., ecological fallacy) *is not* a necessary outcome of *any* data aggregation, and that seemingly unexpected results may often stem from a misguided variable selection. The implications of the results of the analysis for epidemiological studies are discussed, and recommendations for public health policy are formulated.

Keywords: ecological fallacy; data aggregation; air pollution; health effects; pulmonary function tests

1. Introduction

In his seminal paper, Robinson (1950) distinguished between two types of correlation - ecological and individual. The former is obtained for a *group* of people, while the latter is estimated for indivisible units, such as *individuals.* According to Robinson's line of argument, ecological and individual correlations tend to be dissimilar. As a result, any assumption about an individual based on average data obtained for a group to which the individual belongs may result in an assessment error, known as "ecological fallacy" (Elliot et al., 1996; Morgenstern and Thomas, 1993; Rothman, 1986).

Although Robinson's article in American Sociological Review (ibid.) became a real eye-opener for many social scientists, more than a decade earlier, Gehlke and Biehl (1934) reported a similar variation of correlation coefficients in line with data aggregation. Follow up studies (see *inter alia* Openshaw, 1984; Unwin, 1996) shed additional light on Gehlke-Biehl-Robinson's findings, showing that the size of correlation coefficients tends, in general, to increase with data aggregation into areal units of larger size. Openshaw (1984) termed this phenomenon the "modifiable areal unit problem" or MAUP.

The awareness about ecological fallacy has not affected geographic research at any considerable extent, where aggregate data are widely used both for empirical analysis and forecasting (see, for example, Glaeser et al., 1992; Felsenstein and Portnov, 2005). However, in social and epidemiological studies, the situation appears to be different. Due to the "ecological fallacy" concern, the use of aggregate data in these studies has either become a taboo or is being treated with caution (Elliot et al., 1996; Greenland, 2001; Openshaw, 1984).

Under which circumstances does "ecological fallacy" occur and how strongly may it affect empirical findings?

Characteristically, the most striking example Robinson (1950) drew from the U.S. 1930 Census of Population and Housing to substantiate his findings - illiteracy *vs.* percent of foreign born, - had relatively little to do with "ecological fallacy" *per se,* but rather with a biased selection of variables. While at the individual level, foreign immigrants in 1930 were generally less educated than "veteran" Americans, the aggregate data in Robinson's study seemed to indicate otherwise: the correlation between percent illiterate (in a region's total population) and percent foreign-born was found to be negative, *implying* that immigrants were more literate than the "natives." However, if illiteracy rates were estimated for the foreign born (as opposed to the total population of regions, calculated by Robinson), the above spurious correlation between immigrant shares and illiteracy rates would have been avoided.

Despite the importance of ecological fallacy concept for empirical research, relatively few studies dealt with this issue in sufficient depth (see *inter alia* Elliot et al., 1996; Greenland, 2001; Lasserre et al., 2000; Morgenstern and Thomas, 1993). Possible reasons are unavailability of individual data (as opposed to areal aggregates which are more readily available) and privacy considerations (Elliot and Wartenberg, 2004; Greenland, 2001; Lasserre et al., 2000; Morgenstern and Thomas, 1993). In addition, even when individual-level health data are accessible, there is a difficulty to match them with socio-economic variables which are usually aggregated into census-designated statistical areas (i.e., census blocks and tracts), and rarely available at the individual level (Elliot et al., 1996; Elliot and Wartenberg, 2004; Nuckols et al., 2004). Although geographic information systems (GIS) technology, which has become widely available in recent years (Brauer et al., 2003; Elliot and Wartenberg, 2004; Cockings et al., 2004; Nuckols et al., 2004; Scoggins et al., 2004), may simplify the establishment of such data linkages and thus help to verify the correspondence between results obtained from individual data and those obtained from areal aggregates, such comparative studies are yet largely forthcoming.

The present paper attempts to revisit the ecological fallacy concept by testing the correspondence of results obtained at different levels of areal aggregation. The paper is organized as follows. It starts with general discussion of instances in which substantial differences between group and individual correlations may occur. The linkages between estimates of exposure to air pollution and results of children's pulmonary function tests (PFT), available for the Hadera region of Israel, are then used to verify to validity of the research assumptions. Although various levels of data aggregations affected somewhat the strength of relationships between the research variables, main differences in the analytical outcomes appear to have resulted from alternative variable specifications.

2. Individual vs. group correlations: expected relationships

Suppose that residents of a region were tested for a particular health effect, attributed to the presence of a local environmental factor, e.g., air pollution. For the sake of simplicity, let's assume that, prior to the introduction of the environmental factor in question, the individuals covered by the survey did not differ substantially in respect to their health status or any other parameter (viz., welfare, occupation, housing conditions, etc.) that can interfere with the results of the inquiry.

During the first stage of the analysis, the results of the individual tests were mutually compared with individual estimates of exposure to the environmental factor in question and correlation coefficients were calculated. Then, the results of individual tests and individual exposure levels were aggregated into groups, representing regional subdivisions (e.g., townships or small census areas), in which the individuals reside. The correlation analysis of the averages was then rerun.

The question we shall try to answer is as follows: Under which circumstances may the results of the two stages of the analysis (i.e., individual and group correlations) be distinctively different?

To answer this question, let us consider five simplified diagrams shown in Fig. 1. Each large square in Fig.1 represents the region under study, while small squares are internal subdivisions used for cross-area comparison, and grey circles mark areas in which abnormally high concentrations of the environmental factor in question are detected.

<<< Figure 1 about here >>>

The first two cases (Fig. 1a-b) assume that the geographic distribution of the individuals' homes is homogeneous, whereas in three other cases, the individuals are scattered unevenly across the study area (see small black dots in Fig. 1 c-e).

Suppose that ***individual*** correlations, estimated during the first stage of the analysis, were found to be statistically significant. The same (or similar) relationships are likely to emerge under the aggregation scheme featured in Fig. 1a. According to this scheme, the average values of the environmental factor vary by regional subdivision, with the residents of the central part of the region (nearly totally covered by the grey spot) being exposed most.

The outcome of the analysis may, however, be distinctively different if the grouping of individuals follows the regional subdivision scheme diagrammed in Fig. 1b. According to this scheme, the individuals covered by the study spread ***evenly*** across the region, and all regional subdivisions are ***equally exposed*** to the environmental factor in question. Although individuals near the center are more exposed than elsewhere (Fig. 1b), the shares of the exposed are equal in each subdivision. As a result, neither significant difference in the observed health effects across regional subdivisions can be detected at the group (subdivision) level, nor significant correlation of aggregated data may occur, resulting in the case of ecological fallacy ***par excellánce.***

The skewed distribution of individuals (Fig. 1c-e) ***does not*** necessarily alter the outcome of the analysis. Thus, in two cases shown in Fig. 1c and 1d, a cross-group comparison may lead to the same outcome as above. In particular, no significant difference may be found between the four regional subdivisions covered by the study, due to the fact that the individuals in each sub-region are more or less equally exposed to the environmental factor in question, and no significant differences may thus be detected by a *mutual comparison of the averages*.

The situation shown in the last diagram (Fig. 1e) is, however, dissimilar. Due to differences in the distribution patterns of the individuals across the regional subdivisions (Fig. 1e) and their different exposure levels to the environmental factor in question, the relationships observed at the *individual* level are likely to emerge at the *aggregated* level as well.

Summing up, we may conclude that differences between individual and group correlations (i.e., ecological fallacy) ***are not*** a necessary outcome of ***any*** data aggregation and that the situations, in which such differences are likely to occur, may be detectable from the outset of the analysis. In the following sections, we shall attempt to validate this assumption using a case study which permits various levels of data aggregation.

3. Hadera region, Israel as a case study

In 1996-1999, the research staff of the Institute for Environmental Research, Israel Ministry of the Environment repeatedly tested a sample of 1,810 schoolchildren of the 2nd, 5th, and 8th grades from elementary schools in the Hadera district (see Fig. 2). The children covered by the study underwent pulmonary function (PF) tests and their parents filled out detailed questionnaires on their socio-demographic and household characteristics (see Appendix 1: Pulmonary function data).

Accurate street addresses that could be mapped were available for 1,492 out of the 1,810 children (82.4%). This cohort (1,492 children) thus formed the basis for the present investigation, for which information on the location of the children's homes was essential for estimating the concentrations of air pollutants at the places where the children reside. The analysis indicated that the final sample (1,492) was fairly representative of the entire cohort of the relevant school grades who resided in the study area, and in respect of gender and proportion of children in each grade (P<0.05).

During the initial investigation (Barchana et al., unpublished data; Goren et al., unpublished data), the results of the PF tests for 1996 and 1999 were analyzed separately, using four townships - *Hadera-central, Givat Olga, Beit Eliezer, and Pardes-Hanna,* - for data aggregation and analysis (see Fig. 2).

Contrary to initial expectations, the research indicated *no* significant differences in the average values of the PF tests across the townships, despite their distinctively different air pollution levels and the wealth of evidence accumulated to date on the link between air pollution and PF development in children (see *inter alia* Gauderman et al., 2000, 2004; Peters et al., 1999a, 1999b; Pikhart et al., 2000; Schwartz, 2004).

A possible reason for these inconclusive results was the use of *aggregated data* for four relatively large townships, presumably leading to "ecological fallacy" in research results (Dubnov *et al.,* 2006). The rationale for this conclusion is as follows: Since large geographic areas for which composite data are used tend to exhibit considerable intra-regional variations in the local levels of air pollutants, individual exposure levels cannot presumably be inferred from aggregated data, and the outcome is insensitive exposure estimates (Greenland, 2001; Elliot et al., 1996, 2004; Nuckols et al., 2004; Rothman, 1986, 1993).

The immediate goal of the following analysis is to revisit the results of the initial investigation of the Hadera data, in an attempt to determine whether the inconclusive results reported by the initial inquiry (viz., Barchana et al., unpublished data; Goren et al., unpublished data) were indeed attributed to ecological fallacy or some other underlying causes.

4. Data processing and analysis

4.1. Research phases

The data were processed and analyzed in the following *two* phases.

- During the ***first phase,*** the results of individual PF tests and air pollution estimates (see Appendix 1) were aggregated into ***two sets*** of areal units: a) four townships identical to those used in Barchana ***et al.*** (unpublished data), and b) 20 small census areas (SCAs), similar in size to Census Block Groups in the U.S.A. (see Fig. 2 and Appendix 2).
- During the ***second phase***, the homes of the children participating in the survey were positioned on the map (see small grey dots in Fig. 2) and the levels of air pollution to which each child was exposed were calculated. The task was performed in the ArcGIS 9^{TM} software, using its "spatial join" tool (Minami, 2000). Individual exposure levels were calculated separately for NOx and SO_2 (see Appendix 1: Air pollution data).

<<< Figure 2 about here >>>

4.2. Data analysis

The statistical analysis was performed in two steps. First, the average PF rates observed in 4 townships and 20 small census areas (SCAs) at the end of the study period (i.e., in 1999 - see Appendix 1: Pulmonary function data) were juxtaposed with ***average*** levels of air pollution detected in these areas. Then, the results of ***individual*** PF tests, and of ***individual*** pollution estimates were mutually compared with demographic and health variables, potentially affecting pulmonary function, viz.: child's age and height at the start of the study period (i.e., in 1996); gender; presence (or absence) of pulmonary diseases diagnosed by a doctor; overall duration of residence in the study area; exposure to environmental tobacco smoking in the family; housing density; education levels of both parents; and proximity to main roads to control for exposure to air pollution from motor vehicles. The confounding role of these variables has been outlined by most previous studies (Gauderman et al., 2004; Goren et al., 1991; Goren and Hellmann, 1995; Jedrychowski et al., 2002; Peters et al., 1999a, 1999b; Peled et al., 2001).

In the initial phase of the analysis, the Kruskal-Wallis test and bivariate correlation analysis were run to determine respectively the significance of inter-group differences, and the correlation strength between average pollution levels and PFT data. Next, the Multiple Regression Analysis (MRA) was used, to identify and measure the effects of the aforementioned explanatory variables on the individual PFT values. During the analysis, spatial autocollinearity of residuals (Moran's *I* diagnostic), multicollinearity, normality, and homogeneity of variance assumptions were tested and their results were found satisfactory.

During the analysis, both logarithmic and exponential transformations of the NOx and SO2 variables (see Appendix 1: Air pollution data) were tested. From the outset the relationship between PFT values and air pollution levels was presumed to be non-linear. For instance, we expected a disproportionably greater damage to occur under higher concentrations of air pollutants than under moderate and low concentrations. This non-linearity of relationship appeared to be captured best by exponential transformations. In the following discussion only the best performing models (for the NOx exponent) are reported.

As Fig. 2 shows, the children covered by the sample are distributed unevenly both across the study area and its internal subdivisions (townships). As a result, they appear to have different long-term exposures to air pollution, with the children living in ***Pardes-Hanna*** being exposed most (see small grey dots and NOx contour lines in Fig. 2). The distribution map thus closely resembles the situation featured in Fig. 1e, according to which no ecological fallacy in data interpretation should expectedly occur.

4.3. "Hot spot" analysis

Indicators of spatial association (such as, Moran's I, Geary's C, Gi(d), and Gi*(d)) provide summary information about the intensity of spatial interaction between values observed in adjacent locations, thus helping to determine whether a parameter's values are arranged in space in a systematic manner. Such a systematic distribution of values is known as "spatial autocollinearity" or "spatial association" (Cliff and Ord, 1981; Anselin, 1999).

In the present study, the analysis of *local spatial autocorrelation* was used to detect 'hot spots' in the spatial distribution of PFT values (see Appendix 3).

5. Results

As Table 1 shows, at *neither* level of aggregation (townships, SCAs, and individuals), the bivariate correlation between FVC_99p (*forced vital capacity in 1999* - for more details see Appendix 1: Pulmonary function data) and NOx estimates appears to be statistically significant ($P>0.05$). With the exemption of the second level of aggregation, at which the correlation between average PFT values and air pollution estimates is marginally significant ($P=0.045$), the results for FEV1_99P are similar (see Table 1).

Although the strength of NOx-PFT relationship tends to increase initially with data desegregation (FVC_99p: $P<0.1$ for SCA *vs.* $P>0.8$ for townships; Table 1), it appears to drop with further desegregation (i.e., FVC_99p: $P>0.3$ for individuals).

Characteristically, the mean values of FEV1_99P and FVC_99P do not differ significantly across either townships or SCAs (Table 2), which is, generally, in line with the results of the initial investigation (Barchana et al., unpublished data).

The air pollution variable (NOx estimate) does not emerge as statistically significant in the multiple regression analysis either, in which this factor is controlled for the effect of potential confounders, such as: welfare, environmental tobacco smoking, etc. (Table 3).

<<< Tables 1-3 about here >>>

The absence of any clear aggregation-induced trend suggests that the lack of significant relationship between air pollution estimates and PFT values *cannot* thus be attributed to data aggregation and averaging, as might be expected.

Another possible explanation is that the PFT variables used in the analysis until now – FVC_99p and FEV1_99p, – may be somewhat problematic, as illustrated by way of a hypothetical example featured in Table 4.

Suppose that the study area is subdivided into four sub-areas (1-4), which exhibit different air pollution levels (Table 4). In ***year**$_1$*, individuals residing in these sub-areas underwent PF tests and the average values of these tests indicated significant differences (PF_{year1}: P<0.05; Table 4). In ***year**$_{1+n}$*, PFT was rerun ($PFT_{year1+n}$; Table 4). Although changes in PFT between ***year**$_1$* and ***year**$_{1+n}$* are negatively correlated to air pollution levels (ΔPFT: r=-0.970; P<0.05), there are ***no*** differences whatsoever in the average values of the repeat test ($PF_{year1+n}$=95%; Table 4). Therefore, using the latter numbers for a cross-sectional comparison may result in the absence of any clear relationship.

<<< Table 4 about here >>>

Although the case featured in Table 4 is clearly hypothetical, it may reflect, at least in theory, what happened in the study area under investigation. To verify this assumption, we rerun our analysis using ΔPFT (ΔFVC and ΔFEV1) as dependent variables. The results of the repeat analysis are reported in Tables 5 – 7.

<<< Tables 5 - 7 about here >>>

As expected, the outcome the analysis has changed. In particular, the bivariate correlation between ΔPFT (ΔFVC and ΔFEV1) and NOx increased to -0.887/ -0.902 for townships (P<0.10; Table 5), as compared to -0.124/0.035; P>0.8, obtained in the initial analysis (see Table 1). Similar changes are observed at other resolution levels as well: r=-0.770/-0.765 for SCAs; P<0.01; Table 5 (as opposed to r=-0.403/-0.452; P>0.04; Table 1), and r=-0.133/-0.124 for individuals; P<0.001; Table 5 (as compared to -0.027/-0.034; P>0.1; Table 1).

Notably, at *all* aggregation levels (townships, SCA, and individuals), the ΔPFT-NOx relationship exhibits the expected negative sign and even gains its statistical significance with areal disaggregation ($P>0.05$ for townships, vs. $P<0.0001$ for individuals; see Table 5).

The differences between sub-areas (both townships and SCAs) also became highly significant ($P<0.001$; Table 5), and the air pollution factor emerged as statistically significant in the regression models estimated for both SCAs (B=-138.541; $P<0.01$ for ΔFVC; B=133.515; $P<0.01$ for ΔFEV1; Table 6) and individuals (B=-121.143 for ΔFVC, and B=-119.203; $P<0.01$ for ΔFEV1; Table 7).

Other factors evaluated for possible effects of covariates on ΔPFT are: road proximity, length of child's residence in the study area, housing density, level of father's education, child's gender, passive smoking in the family, and presence of pulmonary diseases. In addition to air pollution, only age and height of a child at the start of the study period appears to be statistically significant in the models estimated for individuals ($P<0.05$; see Table 7). [In theory, age and height are collinear variables. However, our multicollinearity test found that their actual correlation for the sample data was within tolerable limits (Tol.>0.3), and unlikely to cause a significant bias of regression estimates].

The road proximity is also marginally significant in the ΔFVC model (t=-1.724; $P<0.1$; Table 7). As expected, the sign of this variable is negative (B= -1.165), implying that proximity to a major road tended to reduce, *ceteris paribus,* a child's ΔFVC by some 1.2%.

5.1. Local spatial autocorrelation patterns

The results of local autocorrelation analysis for the ΔFVC variable are reported in Fig. 3. [The results for ΔFEV1 appear to be similar and are not reported here for the sake of brevity]. In this diagram, large red dots indicate *clusters of children* with significantly low values of ΔFVC (compared to the global mean), while large blue dots show clusters of children whose

FVC change in 1996-1999 was significantly ***higher*** that the average value for the study cohort as a whole.

<<< Figure 3 about here >>>

As Fig. 3 shows, there is a clear clustering of significantly ***low*** ΔFVC values in the ***Pardes-Hanna*** township (see red dots in the upper right corner of the map). This township is primarily affected by air pollution from the power station, as indicated by extremely high levels of the NOx contours (NOx estimates > 30 ppm). Concurrently, there are also a few clusters of negative ΔFVC values elsewhere, specifically in the ***Hadera- central*** township (center of the map). This township is generally characterized by moderate levels of station-generated NOx pollution (NOx estimate < 30 ppm). A more detailed investigation of the location of these clusters points out at their close proximity to a major road (old road "Tel Aviv-Haifa"), thus making the clustering of negative ΔFVC values in this part of the study area quite explicable.

6. Discussion

In the present study, the strength of the linkages between children's PFT values and air pollution estimates was tested for three levels of areal aggregation – 4 townships, 20 small census areas (SCAs), and 1492 individuals covered by the survey. During the analysis, two separate sets of measures were used: two cross-sectional estimates of children's PFT performance at the end of the study period (***FVC_99p*** and ***FEV1_99p***) and two longitudinal estimates of PFT change between 1996 and 1999 (ΔFVC and ΔFEV1). The analysis was performed using four different statistical techniques: bivariate correlation analysis; a non-parametric test of intergroup differences (the Kruskal-Wallis test); the analysis of local spatial autocorrelation (the Getis-Ord "hot spot" index), and multivariate regression analysis. In the latter models, children's exposure to air pollution was adjusted for children's demographic and health characteristics, such as education of parents, housing conditions, presence of pulmonary diseases, etc.

The association between children's pulmonary function (PF) development and their exposure to air pollution was investigated in two phases. During the first phase, PFT *averages* were mutually compared with *average* levels of air pollution estimated for the statistical areas in which the children reside. During the second phase of the analysis, *individual* pollution estimates were compared with *individual* PFT results.

Since the same group of children, tested in 1996 and again in 1999, was used in the analysis, and their health or household characteristics did not change dramatically over the study period, no selection bias was present in the data. We also compared mapped and unmapped participants for distribution of mean age and gender. No significant differences in these distributions were found, so we could conclude that no bias in mapping might possibly account for our findings.

While different levels of areal data aggregation (townships-SCAs-individuals) changed the outcome of analysis only marginally, the selection of indices measuring the children's PF performance had a significant influence on the outcome of the analysis. In particular, while the use of cross-sectional data (PFT values at the end of the study period) failed to detect any significant link between air-pollution estimates and children's PF performance, the use of PFT change indices (ΔFVC and ΔFEV1) indicated substantial differences between groups (sub-areas) and strong negative correlations of ΔPFT with air pollution estimates, at any level of areal aggregation we tested.

Characteristically, the statistical significance of correlation between ΔPFT and pollution estimates appeared to increase with data disaggregation, from $P>0.05$ for townships to $P<0.01$ for individuals, thus implying that the *relationship detected at the aggregated data level are likely to emerge at the individual level as well, or even become stronger.*

7. Conclusions

Ecological fallacy is hardly an imaginary phenomenon. As we argue, under certain circumstances, the areal aggregation of data may indeed lead to erroneous estimates. For instance, when the residents of *each* geographic sub-region under study are *equally* affected by an environmental factor in question (see Fig. 1b-d), then, indeed, the aggregation of data may lead to erroneous estimates. However, in other cases (e.g., substantially different levels of exposure across geographic sub-areas, e.g., see Fig.1a,e), *no* ecological fallacy in data interpretation should arguably occur, and the linkages identified for areal aggregates are likely to emerge at the individual level as well.

In general, the possibility of ecological fallacy may be detected by scrutinizing the distribution map, as the present study demonstrates. Thus, the juxtaposition of the location of the children's homes in our survey with the boundaries of geographic subdivisions, and the patterns of air-pollution in the area (Fig. 2) indicated that ecological fallacy was unlikely. Indeed, as our analysis demonstrated, the inconclusive results about the links between air pollution levels and health effects, obtained by the initial investigation, were attributed to a misguided variable selection (i.e., using cross-sectional values of PFT instead of more sensitive ΔPFT), rather to ecological fallacy *per se*, as could be expected.

There is thus no *a priory* reason for avoiding the use of aggregate data in epidemiological studies. Aggregated data are often more readily available for researchers than individual estimates; they are easy to process, analyze and link to other information sources (such population enumerations), and may give nevertheless an accurate indication about the relationships which may expectedly be found in follow-up investigations, or necessitate such in-depth investigations, if necessary (Elliot and Wartenberg, 2004). The essential condition is, however, the ability of the researcher to identify from the outset of the analysis the situations in which ecological fallacy is likely to occur and thereby interfere with the results of the investigation.

In our view, this task may be performed by the map analysis tools provided by geographic information systems, which gain popularity in recent years.

An obvious limitation of the present study is unaccountability for particulate matter of less than 2.5 μm or less than 10 μm in aerodynamic diameter (i.e., $PM_{2.5}$ and PM_{10}), especially in light of recent publications demonstrating the adverse effects of this air pollutant on different aspects of human health (Pope et al., 2002, 2004; Samet et al., 2000; Schwartz, 2004; Ward and Ayres, 2004). However, as we believe, this limitation does not influence significantly the results of our study, which primary goal was to identify changes in analytical outcomes attributed to different levels of data aggregation, rather then demonstrating the influence of a particular air pollutant on public health. A more detailed investigation of such effects may represent a legitimate topic for future studies.

Finally, we should note that the empirical results of the present study are definitely location specific. Follow up studies carried out elsewhere may thus be needed for the verification of generality of our empirical results.

References

American Thorasic Society. 1995. Standardization of spirometry, 1994 update. Am J Respir Crit Care Med, Sep, 152(3): 1107–1136.

Anselin L. 1999. Spatial Econometrics. Dallas, TX: Bruton Center, School of Social Sciences, University of Texas.

Association of Towns for Environmental Protection. 2005. Annual Report for 2004. Hadera, Israel [in Hebrew] (Available at: http://www.igudhadera.co.il/).

Brauer M., Hoek G., van Vliet P., Meliefste K., Fischer P., Gehring U., et al. 2003. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. Epidemiology, Mar, 14(2): 228–239.

Cliff A.D. and Ord J.K. (1981) Spatial processes – Models and Applications, Pion, London.

Cockings S., Dunn C.E., Bhopal R.S., and Walker D.R. 2004. Users' perspectives on epidemiological, GIS and point pattern approaches to analyzing environment and health data. Health Place, June, 10(2): 169–182.

Dubnov, J. Barchana M. Rishpon S., Leventhal A., Segal I., Carel R. and Portnov B.A. 2006. Estimating the effect of air pollution from a coal-fired power station on the development of children's pulmonary function, Environ Research (in press).

Elliot P. and Wartenberg D. 2004. Spatial epidemiology: current approaches and future challenges. Environ Health Perspect. Jun;112(9):998-1006. Review.

Elliot P., Cuzick J., English D., and Stern R. 1996. Geographical and Environmental Epidemiology. Methods for Small Area Studies. WHO Europe Regional Office, Oxford University Press.

Enright P.L., Linn W.S., Avol E.L., Margolis H.G., Gong H. Jr., Peters J.M. 2000. Quality of Spirometry Test Performance in Children and Adolescents Experience in a Large Field Study. Chest; 118:665-671

Felsenstein D. & Portnov B.A. (Eds.) 2005. Regional Disparities in Small Countries. Springer Verlag, Heidelberg etc.

Feris B.G. 1978. Recommended respiratory disease questionnaire for use with adults and children in epidemiological research. Epidemiology standardization project. Am. Rev. Respir Dis; 118: 1–120.

Gauderman W.J., McConnell R., Gilliland F., London S., Thomas D., Avol E.L., et al. 2000. Association between air pollution and lung function growth in Southern California children. Am J Respir Crit Care Med, 162: 1383–1390.

Gauderman W.J., Avol E.L., Gilliland F., Vora H., Thomas D., Berhane K., et al. 2004. The effect of air pollution on lung development from 10 to 18 years of age. N Engl J Med, 351: 1057– 1067.

Gehlke, C. and Biehl, K., 1934. Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material, J of Am Statistical Association, 29:169-170.

Glaeser E.L., Kallal H.D., Scheinkman J.A. and A. Shleifer (1992) Growth in Cities, The Journal of Political Economy, 100(6): 1126-52.

Goren A.I., Goldsmith J.R., Hellmann S., Brenner S. 1991. Follow-up of schoolchildren in the vicinity of a coal-fired power plant in Israel. Environ Health Perspect, Aug; 94: 101–105.

Goren A.I., Hellmann S. 1995. Respiratory conditions among schoolchildren and their relationship to environmental tobacco smoke and other combustion products. Arch Environ Health, Mar-Apr; 50(2): 112–118.

Goren A.I., Hellmann S., Glaser E.D. 1995. Use of outpatient clinics as a health indicator for communities around a coal-fired power plant. Environ Health Perspect, Dec, 103(12): 1110–1115.

Greenland S. 2001. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. Int J Epidemiol. Dec; 30(6):1343-50

Hankinson J., Odencrantz J., Fedan K. 1999. Spirometric reference values from a sample of the general U.S. population. Am J Respir Crit Care Med, 159: 179–187.

Jedrychowski W., Maugeri U., Jedrychowska-Bianchi I. 2002. Body growth rate in preadolescent children and outdoor air quality. Environ Res, Sep, 90(1): 12–20.

Lasserre V., Guihenneuc-Jouyaux C., Richardson S. 2000. Biases in ecological studies: utility of including within-area distribution of confounders. Stat Med. Jan 15;19(1):45-59

Minami M. and Environmental Systems Research Institute. 2000. Using ArcMap: GIS by ESRI. Redlands, California, ESRI, pp. 365–392.

Morgenstern H. and Thomas D. Principles of Study Design in Environmental Epidemiology. 1993. Environ Health Perspect. Dec;101 Suppl 4:23-38

Nuckols J.R., Ward M.H., and Jarup L. 2004. Using Geographic Information Systems for Exposure Assessment in Environmental Epidemiology Studies. Environ Health Perspect, June, 112(9): 1007–1015.

Openshaw S. 1984. The modifiable areal unit problem, In: Concepts and Techniques in Modern Geography Monograph Series, 38, 41 pp.

Peled R., Bibi H., Pope C.A. 3rd, Nir P., Shiachi R., Scharff S. 2001. Differences in lung function among school children in communities in Israel. Arch Environ Health, Jan-Feb, 56(1): 89–95.

Peters J.M., Avol E.L., Gauderman W.J., Linn W.S., Navidi W., London S.J., et al. 1999. A study of twelve southern California communities with differing levels and types of air pollution: II. Effects on pulmonary function. Am J Respir Crit Care Med, 159: 768–775.

Peters J.M., Avol E.L., Navidi W., London S.J., Gauderman W.J., Lurmann F., et al. 1999. A study of twelve Southern California communities with differing levels and types of air pollution: I. Prevalence of respiratory morbidity. Am J Respir Crit Care Med, 159: 760–767.

Pikhart H., Bobak M., Kriz B., Danova J., Celko M.A., Prikazsky V., et al. 2000. Outdoor air concentrations of nitrogen dioxide and sulfur dioxide and prevalence of wheezing in school children. Epidemiology, Mar; 11(2): 153–160.

Pope C.A. 3rd, Burnett R.T., Thun M.J., Calle E.E., Krewski D., Kazuhiko I., et al. 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA, 287: 1132–1141.

Pope C.A. 3rd, Burnett R.T., Thurston G.D., Thun M.J., Calle E.E., Krewski D., et al. 2004.Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of diseases. Circulation, 109: 71–77

Enright P.L., Linn W.S., Avol E.L., Margolis H.G., Gong H., Peters J.M. 2000. Quality of spirometry test performance in children and adolescents : experience in a large field study. Chest. Sep;118(3):665-71.

Robinson W.S. 1950. Ecological correlations and the behavior of individuals, Am. Sociological Review, 15, 351-57.

Rothman K.J 1986. Modern Epidemiology. Boston: Little, Brown and Co.; pp. 74, 304.

Rothman K.J. 1993. Methodological Frontiers in Environmental Epidemiology. Environ Health Perspect, Dec; 101 Suppl 4:19-21.

Samet J.M., Dominici F., Curriero F.C., Coursac I., Zeger S. 2000. Fine particulate air pollution and mortality in 20 U.S. cities, 1987-1994. N Engl J Med, 343: 1742–1749.

Schwartz J. 2004. Air pollution and children's health. Pediatrics, Apr; 113(4 Suppl): 1037–1043.

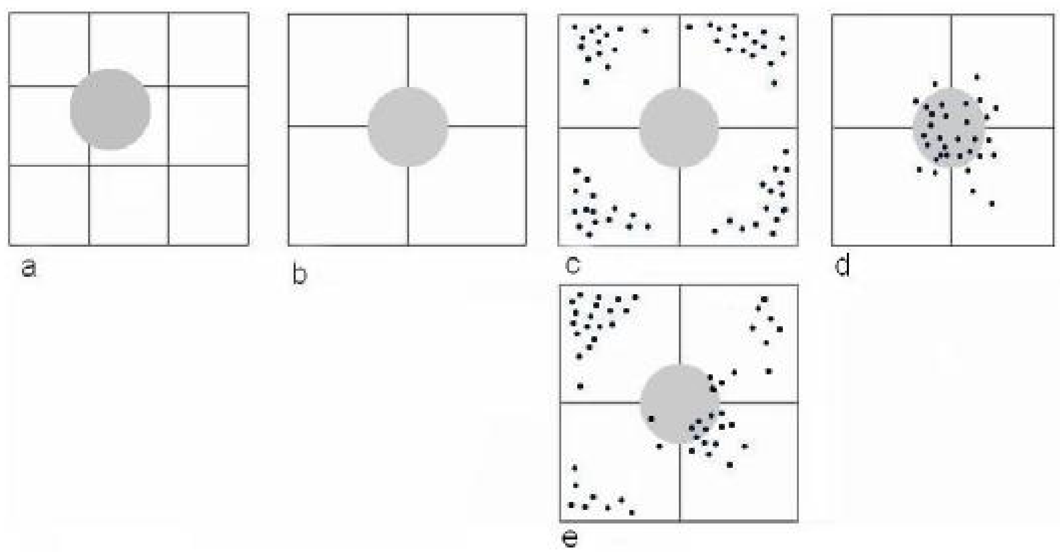
Scoggins A., Kjellstrom T., Fisher G., Connor J., Gimson N. 2004. Spatial analysis of annual air pollution exposure and mortality. Sci Total Environ, Apr 5; 321(1-3): 71–85

Unwin D. J. 1996. GIS, spatial analysis and spatial statistics', Progress in Human Geography 20(4): 540-441.

Ward D.J. and Ayres J.G. 2004. Particulate air pollution and panel studies in children: a systematic review. Occup Environ Med. Apr; 61(4): e13.

EXHIBITS

Figure 1
Typical situations leading (b-d) and not leading (a, e) to "ecological fallacy" upon data aggregation



Note: See text for explanations

Figure 2
Study Area

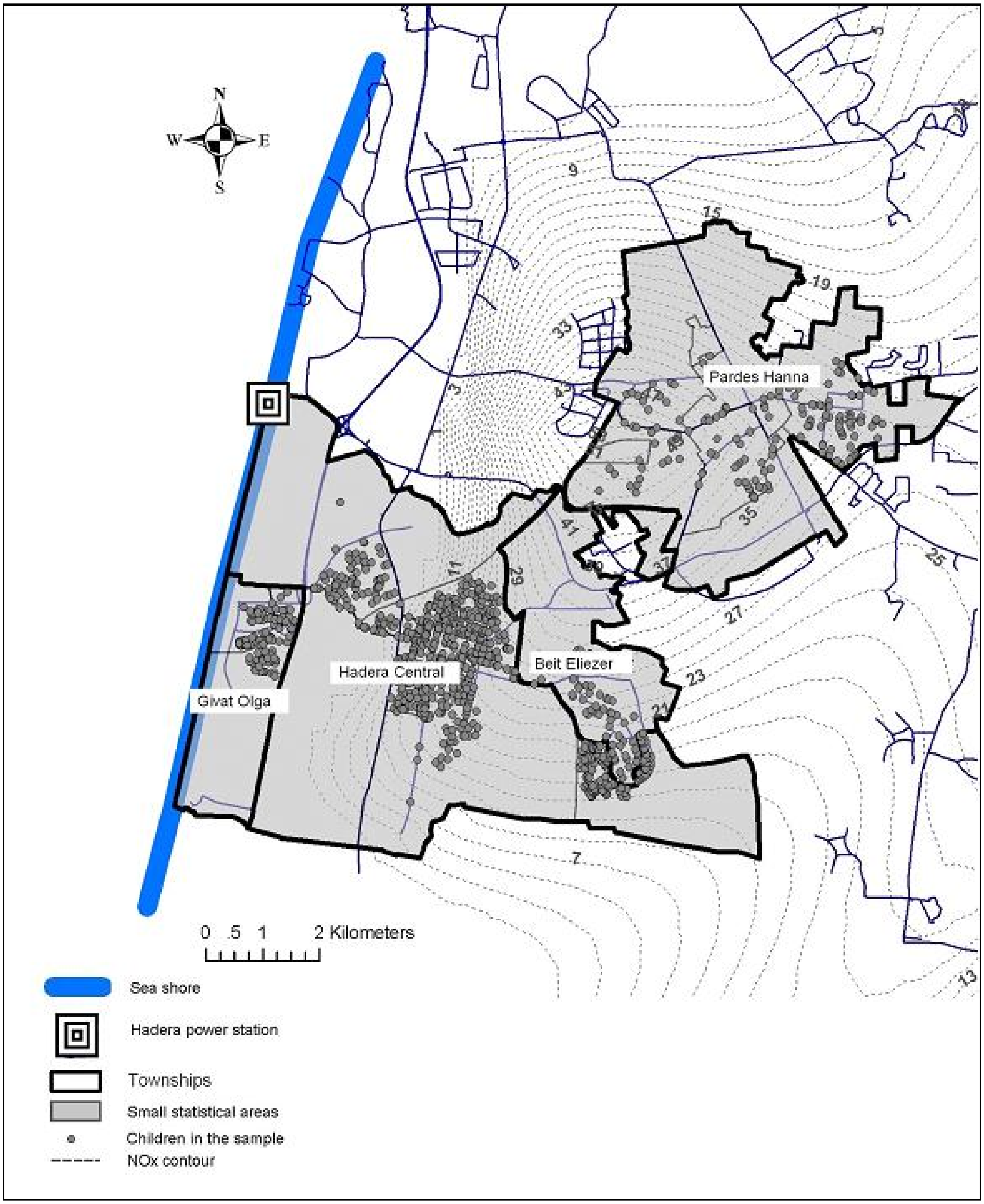


Figure 3
"Hot Spot" Analysis of ΔFVC values (Gettis-Ord Statistic)

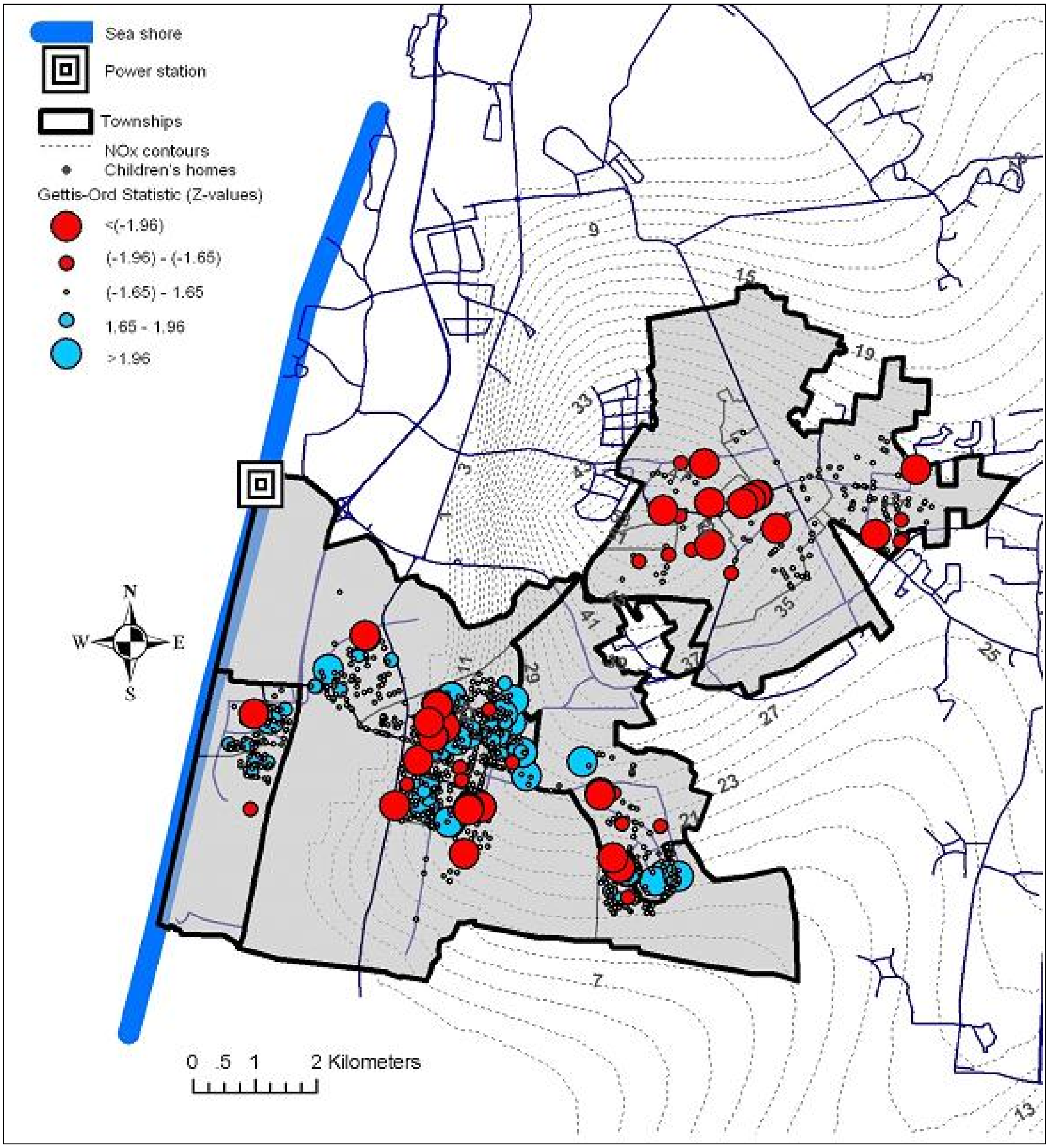


Table 1
Bivariate coefficients of correlation and results of Kruskal-Wallis Test for different levels of data aggregation

Variable	Bivariate correlation between PFT values and NOx estimates at different levels of data aggregation			Kruskal-Wallis test for significance of inter-area differences (χ^2 values)	
	Townships	Small Census Areas (SCA)	Individuals	Townships	Small Census Areas (SCA)
FVC_99p	-0.124	-0.403	-0.027	5.212	28.939
	(0.876)	(0.078)	(0.302)	(0.157)	(0.067)
FEV1_99p	0.035	-0.452	-0.034	2.285	31.993
	(0.965)	(0.045)	(0.185)	(0.515)	(0.031)
N (df)	4	20	1492	3	19

Note: Actual significance levels are in parentheses

Table 2

Factors affecting PFT values (units of analysis – small census areas; method - multiple regression; dependent variables – FVC_99p and FEV1_99p)

Variable	FVC99_p			FEV1_99p			VIF[d]
	B[a]	T[b]	Sig.[c]	B[a]	T[b]	Sig.[c]	
(Constant)	106.912	2.263	0.038	134.658	2.671	0.017	
NOx level [e]	-34.669	-1.057	0.306	-50.764	-1.450	0.166	1.388
Income	0.569	0.455	0.655	0.397	0.297	0.770	1.018
Age	2.244	1.011	0.327	1.444	0.609	0.551	1.401
R^2	0.232			0.232			
R^2 Adjusted	0.088			0.088			
F	1.611		0.226	1.612		0.226	
No of cases	20			20			
Z Normal I[f]	-0.395			-1.134			

[a] Unstandardized regression coefficient; [b] t-statistic; c actual significance of t-statistic; [d] variance inflation factor (multicollinearity diagnostic); [e] transformed to exponent as follows: NOx = exp (NOx [ppm] $*10^{-3}$); [f] spatial auto-collinearity of residuals diagnostic (median neighbor distance = 1000 m).

Table 3

Factors affecting PFT values (units of analysis – individuals; method - multiple regression; dependent variables – FVC_99p and FEV1_99p)

Explanatory variable	Dependent variable						VIF[d]
	FVC99_p			FEV1_99p			
	B[a]	T[b]	Sig.[c]	B[a]	T[b]	Sig.[c]	
(Constant)	120.300	3.855	0.000	115.466	3.542	0.000	
Age	-0.006	-0.014	0.989	-0.226	-0.527	0.598	3.088
Asthma	-0.767	-0.570	0.569	-3.685	-2.621	0.009	1.126
Bronchitis	0.070	0.072	0.943	0.300	0.295	0.768	1.143
Estimated NOx level[e]	-41.388	-1.374	0.170	-41.532	-1.319	0.187	1.021
Father's education	0.158	1.168	0.243	0.115	0.812	0.417	1.061
Gender	0.652	0.900	0.369	1.236	1.633	0.103	1.012
Height	0.126	2.196	0.028	0.202	3.366	0.001	2.882
Housing density	-0.798	-1.237	0.216	-0.390	-0.579	0.563	1.039
Passive smoking	-0.028	-0.038	0.970	-0.271	-0.354	0.723	1.036
Proximity to main road[g]	-0.818	-1.079	0.281	-0.749	-0.946	0.344	1.021
Years in study area	-0.332	-2.095	0.036	-0.356	-2.151	0.032	1.318
R^2	0.022			0.039			
R^2-adjusted	0.010			0.028			
Number of cases	1492			1492			
F statistic	1.883	0.038		3.443	0.000		
Z-Moran's I[5]	-0.684			0.494			

[a] Unstandardized regression coefficient; [b] *t*-statistic; c actual significance of *t*-statistic; [d] variance inflation factor (multicollinearity diagnostic); [e] transformed to exponent as follows: NOx = exp (NOx [ppm] $*10^{-3}$); [f] spatial auto-collinearity of residuals diagnostic (median neighbor distance = 40 m); [g] recorded to dichotomous variable, as follows: An area closer than 50 m (the first row of buildings) to a main road's longitudinal axis was conditionally defined as a high exposure area (1); otherwise it was defined as a low exposure area (0).

Table 4
Hypothetical case illustrating differences between PFT and ΔPFT values

Sub-area	Air Pollution Level	PFT_{year1}	ΔPFT	$PFT_{year1+n}$
1	20	89	6.0	95
2	30	90	5.0	95
3	50	100	-5.0	95
4	60	102	-7.0	95

Table 5
Bivariate coefficients of correlation and results of Kruskal-Wallis Test for different levels of data aggregation (ΔPFT values)

Variable	Bivariate correlation between PFT values and NOx estimates at different levels of data aggregation			Kruskal-Wallis test for significance of inter-area differences ($\chi 2$ values)	
	Townships	Small Census Areas (SCA)	Individuals	Townships	Small Census Areas (SCA)
ΔFVC	-0.887	-0.770	-0.133	50.981	87.479
	(0.113)	(<0.001)	(<0.0001)	(<0.001)	(<0.001)
ΔFEV1	-0.902	-0.765	-0.124	41.529	76.524
	(0.098)	(<0.001)	(<0.0001)	(<0.001)	(<0.0001)
N (df)	4	20	1492	3	19

Note: Actual significance levels are in parentheses

Table 6

Factors affecting ΔPFT values (units of analysis – small census areas; method - multiple regression; dependent variables – ΔFVC and ΔFEV1)

Variable	ΔFVC			ΔFEV1			VIF[d]
	B[a]	T[b]	Sig.[c]	B[a]	T[b]	Sig.[c]	
(Constant)	-53.090	-2.255	0.039	-74.638	-3.239	0.005	
Estimated NOx level[e]	-138.541	-3.566	0.003	-133.515	-3.512	0.003	1.420
Income	0.334	0.336	0.741	0.232	0.240	0.814	1.035
Age	5.832	2.305	0.035	7.899	3.190	0.006	1.387
R^2	0.699			0.749			
R^2 Adjusted	0.642			0.702			
F	12.379		0.000	15.928		0.000	
No of cases	20			20.000			
Z Normal I[f]	0.717			0.647			

* See footnote to Table 2

Table 7

Factors affecting ΔPFT values (units of analysis – individuals; method - multiple regression; dependent variables – ΔFVC and ΔFEV1)

Explanatory variable	Dependent variable					
	ΔFVC			ΔFEV1		
	B[a]	T[b]	Sig.[c]	B[a]	T[b]	Sig.[c]
(Constant)	53.673	1.936	0.053	36.125	1.260	0.208
Age	0.735	2.017	0.044	2.653	7.040	<0.001
Asthma	0.139	0.117	0.907	0.117	0.095	0.924
Bronchitis	0.742	0.856	0.392	0.973	1.085	0.278
Estimated NOx level [e]	-121.143	-4.522	<0.001	-119.203	-4.302	<0.001
Father's education	0.129	1.062	0.288	0.169	1.347	0.178
Gender	0.472	0.732	0.464	0.132	0.199	0.842
Height	0.442	8.704	<0.001	0.401	7.634	<0.001
Housing density	-0.221	-0.369	0.712	0.024	0.039	0.969
Passive smoking	0.885	1.362	0.174	1.085	1.614	0.107
Proximity to main road[g]	-1.165	-1.724	0.085	-0.812	-1.162	0.245
Years in study area	-0.038	-0.265	0.791	-0.006	-0.039	0.969
R^2	0.274			0.401		
R^2-adjusted	0.265			0.393		
No of cases	1492			1492		
F statistic	31.277	0.000		55.428	0.000	
Z-Moran's I[f]	1.441			1.684		

* See footnote to Table 3

APPENDIX 1

Description of Data Sources

Pulmonary function data

Spirometry was performed by means of a Minato ® AS 500 spirometer and in compliance with the American Thoracic Society (ATS) criteria. Each child performed 3 consecutive pulmonary function tests (PFT), and the maneuver with the largest sum of Forced Vital Capacity (FVC) and Forced Expiratory Volume during the first second (FEV1) was recorded as a representative test (American Thoracic Society, 1995; Enright et al., 2000). Predicted PFT values were calculated using a polynomial model, separately for each gender (Hankinson et al., 1999).

The calculations were performed separately for differences in *forced vital capacity* (FVC) and *forced expiratory volume during the first second* (FEV1).

Then, the relative changes in pulmonary function tests (Δ*PFT)* from 1996 to 1999 were calculated as follows:

ΔFVCi =FVC_99pi - FVC_96pi = FVC_99io*100/FVC_99ie, - FVC_96io*100/FVC_96ie,

ΔFEV1i =FEV1_99pi - FEV1_96pi = FEV1_99io*100/FEV1_99ie, - FEV1_96io*100/FEV1_96ie,

where:

- FVC_99io, FEV1_99io, FVC_96io, and FEV1_96io and PFT96io are observed forced expiratory flow volumes (FVC or FEV1) of child ***i*** in 1999 and 1996, respectively;
- FVC_99ie, FEV1_99ie, FVC_96ie, and FEV1_96ie and PFT96ie are the calculated (expected) volumes for child ***i*** in the same years;
- FVC_99pi, FVC_96pi, FEV1_99pi and FEV1_96pi are respectively FVC and FEV1 performances of child i (observed vs. expected) in 1999 and 1996, expressed as percentages.

To ensure the suitability of ΔPFT estimates for multivariate modeling, normality of distribution was tested by the Kolmogorov-Smirnov (KS) test, in which the distribution of Δ*PFT* values appeared fairly normal (KS $Z<0.9$; $P>0.4$).

Demographic and health data

The questionnaire used in the study was a validated translation of the questionnaire developed and used by the American Thoracic Society (ATS) and National Heart and Lung Institute (Feris, 1978). It includes questions about the presence or absence of pulmonary diseases diagnosed by a physician (e.g., asthma), household-related characteristics, such as gas or oil house heating, housing density, exposure to passive tobacco smoking, parents' education, and duration of living in the study area. The children's parents completed the questionnaires, with an overall rate of return of 72.4%.

Air pollution data

There are 12 monitoring stations in the study area, which provide continuous (24-hours a day) measurements of air pollution levels. For our analysis we used only the measurements simultaneously exceeding half-an-hour reference levels for NOx and SO_2 (0.125 ppm and 0.070 ppm, respectively). These excess concentrations (or so-called "air pollution events") help to distinguish air pollution "splashes" generated by the power station (contributing ~ 50% of emissions in study area) from air pollution constantly present in the area and attributed to other sources such as motor vehicles (Association of Towns for Environmental Protection, 2005; Goren et al., 1995).

For each "air pollution event" we calculated integrated concentration value (ICV) of NOx and SO_2 by multiplying their average concentrations during the "event" [ppm] by the unit of event's duration (half-an-hour is one unit) and then summarized the results over the entire study period (i.e., 1996 through 1999).

These summary values for the 12 air-monitoring stations were then interpolated by krigging, which furnished contours of equal pollution levels for the entire study area. Using these air pollution contours we estimated the individual exposure levels in the vicinity of the children's residences.

The air pollution estimates used in the analysis did not include particulate matter of less than 2.5 μm or less than 10 μm in aerodynamic diameter (i.e., PM2.5 and PM10) because PM measurements were available for only three out of the 12 monitoring stations distributed sparsely across the study area. Due to this limitation we used NOx and SO_2 air pollutants as proxies for air pollution patterns in the study area. This limitation and its implications are addressed in the discussion section.

APPENDIX 2

Descriptive characteristics of selected research variables

Variable /measure	Level of data aggregation		
	Townships	Small Census Areas	Individuals
FVC_96p [%] Mean (min-max) SD	95.1 (92.9 – 99.3) (3.0)	94.6 (89.3 – 100.5) (3.3)	94.2 (53.9 – 140.2) (12.5)
FEV1_96p [%] Mean (min-max) SD	101.2 (98.8 – 105.9) (3.2)	100.9 (94.9 – 106.8) (3.7)	100.4 (52.0 – 153.1) (13.8)
FVC_99p [%] Mean (min-max) SD	93.9 (93.1 – 95.6) (1.1)	93.3 (88.2 – 96.1) (2.1)	93.8 (61.1 – 145.6) (11.4)
FEV1_99p [%] Mean (min-max) SD	97.5 (96.8 – 98.7) (0.8)	97.0 (91.8 – 101.2) (2.3)	97.5 (64.3 – 144.1) (11.8)
ΔFVC [%] Mean (min-max) SD	- 1.2 (-6.3 – 0.6) (3.4)	- 1.2 (-11.1 – 3.9) (3.9)	- 0.4 (-42.3 – 36.5) (11.3)
ΔFEV1 [%] Mean (min-max) SD	- 3.8 (-8.9 – -1.9) (3.5)	- 3.8 (-14.0 – 1.7) (4.2)	- 2.9 (-45.9 – 38.9) (12.7)
NO_x level [ppm] Mean (min-max) SD	17.6 (1.7 – 37.8) (15.1)	17.6 (0 – 48.3) (15.1)	15.6 (0 – 53.0) (11.8)

APPENDIX 3

Getis-Ord measure of local spatial autocorrelation

The Getis-Ord (G_i*(d)) statistic, used in the present analysis for detecting the spatial clustering of abnormally high and low values of PFT variables, is reported as standard normal *z*-values and is calculated as follows:

$$G_i*(d) = \sum_{j}^{n} w_{ij}(d)(x_j - \overline{x_i}) \Big/ (S_i \sqrt{w_i(n-1-w_i) \big/ n-2}),$$

where ***n*** is the number of observations; ***d*** is the distance band within which locations ***j*** are considered as neighbours of the target location ***i***; x_i is the value observed in location ***i;*** $\overline{x_i} = \frac{1}{n-1}\sum_{j}^{n} x_j$; w_{ij} is a symmetric binary weight matrix, whose elements take value 1 if locations ***i*** and ***j*** are neighbours and 0 otherwise, and $S_i^2 = \frac{1}{n-1}\sum (x_j - \overline{x_i})^2$.

Gi*(d) statistic evaluates each point within a network of sites, and helps to determine the relationship between the values observed around the target point and the global mean (Getis and Ord, 1992). This statistic is easy to interpret: a significant and positive Gi*(d) indicates that location i is surrounded by relatively large values (with respect to the global mean) – 'peak-value clusters', whereas a significant and negative Gi*(d) indicates that location i is surrounded by relatively small values – 'dip-value clusters' (ibid.).