Air Qual Atmos Health (2017) 10:411–420
DOI 10.1007/s11869-016-0436-x

Ecological bias in environmental health studies: the problem of aggregation of multiple data sources

Rakefet Shafran-Nathan[1] · Ilan Levy[1] · Noam Levin[2] · David M. Broday[1]

Received: 27 January 2016 / Accepted: 5 September 2016 / Published online: 9 September 2016

Abstract Ecological bias may result from interactions between variables that are characterized by different spatial and temporal scales. Such an ecological bias, also known as aggregation bias or cross-level-bias, may occur as a result of using coarse environmental information about stressors together with fine (i.e., individual) information on health outcomes. This study examines the assumption that distinct within-area variability of spatial patterns of the risk metrics and confounders may result from artifacts of the aggregation of the underlying data layers, and that this may affect the statistical relationships between them. In particular, we demonstrate the importance of carefully linking information layers with distinct spatial resolutions and show that environmental epidemiology studies are prone to exposure misclassification as a result of statistically linking distinctly averaged spatial data (e.g., exposure metrics, confounders, health indices). Since area-level confounders and exposure metrics, as any other spatial phenomena, have characteristic spatiotemporal scales, it is naively expected that the highest spatial variability of both the SES ranking (confounder) and the NOx concentrations (risk metric) will be obtained when using the finest spatial resolution. However, the highest statistical relationship among the data layers was not obtained at the finest scale. In general, our results suggest that assessments of air quality impacts on health require data at comparable spatial resolutions, since use of data layers of distinct spatial resolutions may alter (mostly weaken) the estimated relationships between environmental stressors and health outcomes.

Keyword Data aggregation · Socioeconomic status (SES) · NOx · Modifiable areal unit problem (MAUP) · Ecological fallacy

Introduction

Spatial misclassification refers to potential errors that result from overlaying spatial variables (e.g., exposure metrics, socio-demographic status, and health outcomes) that are represented at different spatial scales, such as census tract, municipal borders, and city neighborhoods. Spatial misclassification has been recognized as a potentially severe limitation in environmental epidemiology (Sheppard et al., 2012). Oftentimes, health-related studies suffer from limited individual-level data on historical exposures to environmental risk factors, risk modifiers, and susceptibility (Stafford et al., 2008; Goodman et al., 2011). For example, many studies look for associations between individual or group-level health data and exposure at either the geocoded residential address or the neighborhood scale, and their results vary depending on the selection of (or constraints in choosing) polygons to represent the areal unit characteristics (Ryan et al., 2008; Diez Roux and Mair 2010). Aggregation of risk metrics in order to combine exposure matrices (e.g., air pollution index—API), or of potential confounding factors that are presented at different spatial scales, may augment the exposure estimation error and lead to increased misclassification between the factors (Sheppard et al., 2012). In particular, spatial associations might be affected by the aggregation process rather than reflect true patterns between the underlying data layers (Jerrett et al., 2010).

✉ David M. Broday
dbroday@tx.technion.ac.il

[1] Environmental, Water and Agricultural Engineering, Faculty of Civil and Environmental Engineering, Technion, Israel Institute of Technology, Haifa, Israel

[2] Department of Geography, Faculty of Social Sciences, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel

In particular, data used in environmental epidemiology studies are often obtained by averaging of records over neighborhood or the city-boundary polygonal representations, resulting in uniform exposure over each polygon (Gryparis et al., 2009). The assumption that associations observed at the area level hold for individuals residing within this area can lead to the so-called ecological fallacy (Wakefield, 2008; Idrovo, 2011). Although many studies specify this problem as a serious limitation, most of them do not account for the ecological bias that may result from the spatial aggregation of health outcome indices and environmental exposure metrics (Wakefield and Shaddick, 2006). Such an ecological bias, also known as aggregation bias or cross-level-bias, may occur as a result of using coarse environmental information about stressors together with fine (i.e., individual) information on health outcomes (Shaddick et al., 2013).

Moreover, one of the challenges in such studies is how to account for the fact that environmental processes occur at varying spatial and temporal scales, and that for a reliable statistics large datasets are needed and large regions normally comprise the study areas. This is specifically true for air pollution and was suggested as one of the reasons for the relatively small contribution of air pollution to the overall burden of disease (Sheppard et al., 2012; Hoskins et al., 2016). Different spatial patterns may emerge as a result of "mapping" the pollutants at different spatial resolutions (Fekete et al., 2010) over the study area, with different pollutants showing different spatial patterns (Yuval and Broday, 2006) due, in part, to their distinct emission height (Amster et al., 2014).

The bias associated with aggregation of multiple spatial data sources (layers), each using distinct geographic units (polygons), is also found in ecology and is termed the modifiable areal unit problem (MAUP; Openshaw 1994; Maantay, 2007; Goodman et al., 2011). The problem emerges when different spatial traits, such as resolution and extent (scale), are used in the same model (Stafford et al., 2008; Flowerdew et al., 2008; Parenteau and Sawada, 2011; Cyril et al. 2013). One aspect of the MAUP is the selection of an optimal geographic unit over which the aggregation of multiple data layers should be performed. Using non-optimal geographic units may increase the spatial misclassification and enhance the so-called boundaries effect. For example, tampering with the size and shape of the geographical units while keeping the total number of units constant can lead to inconsistent across-layer analysis (Flowerdew et al., 2008; Stafford et al., 2008; Poortinga et al., 2008; Parenteau and Sawada, 2011).

The spatial resolution of the data (i.e., the finest cell size or the smallest polygon size) determines the maximum spatial generalization of the information (Stafford et al., 2008; Poortinga et al., 2008). Two opposing processes may occur while aggregating multiple-scale data sources, such as exposure metrics and health outcome indices, into one product (e.g., a risk map). Namely, the associations between the environmental stressors and the health outcomes may (a) improve due to increasing the number of "cases" or due to geographic clustering of the data (e.g., demographic attributes may be available only at the census tract level; Goodman et al., 2011; Lovasi et al., 2008; Zou et al., 2009), or (b) weaken due to loss of accuracy and reduced variability at the coarser scales. Hence, selecting a particular aggregation method over the others may alter the sought associations. In fact, the possibility that a risk factor will introduce ecological bias without having any individual-level association to the studied outcome variable has been noted (Jerrett et al., 2010; Wakefield, 2008; Greenland and Morgenstern, 1989), suggesting that studies in environmental epidemiology may be affected by multi-scale interactions.

To further study this phenomenon, we applied in this study different spatial aggregations to a common exposure metric (ambient concentrations of nitrogen oxides, NOx) and to a potential confounder (the socioeconomic status rank, SES) and evaluated the effect of aggregation on the statistical interaction between them. As explained above, ecological bias arises in environmental health studies from the inability of ecological data to characterize the within-area variability in exposures and confounders (Wakefield, 2008). Here, we demonstrate that distinct within-area variability of spatial risk metric patterns may result from artifacts of the aggregation of the underlying data layers and suggest methods for reducing the ecological bias that results from such aggregation processes. To explore how a change in the spatial resolution and/or the polygonal representation of the data layers affect the interaction between these data layers, we examine possible confounding associations between nitrogen oxide concentrations—a common traffic-related air pollutant—and the SES of the exposed population. Specifically, we demonstrate for the Tel Aviv metropolitan area, Israel, (ca. 3.7 million residents) how different SES polygonal representations and NOx patterns at different spatial resolutions affect the statistical interaction between these two data layers.

Methods

Study area

The study area comprises the Tel Aviv metropolitan area, Israel, extending from the city of Hadera in the north to the city of Ashdod in the south (Fig. 1). The study area is located along the Mediterranean coast, stretching to about 80 km from south to north and 25 km from west to east. The area is characterized by high density of roads and railways, two airports, and one commercial seaport. The total population in this area is about 3.7 million inhabitants that live in 57 municipalities and 768 census blocks, defined by the Israel Central Bureau of Statistics (CBS).

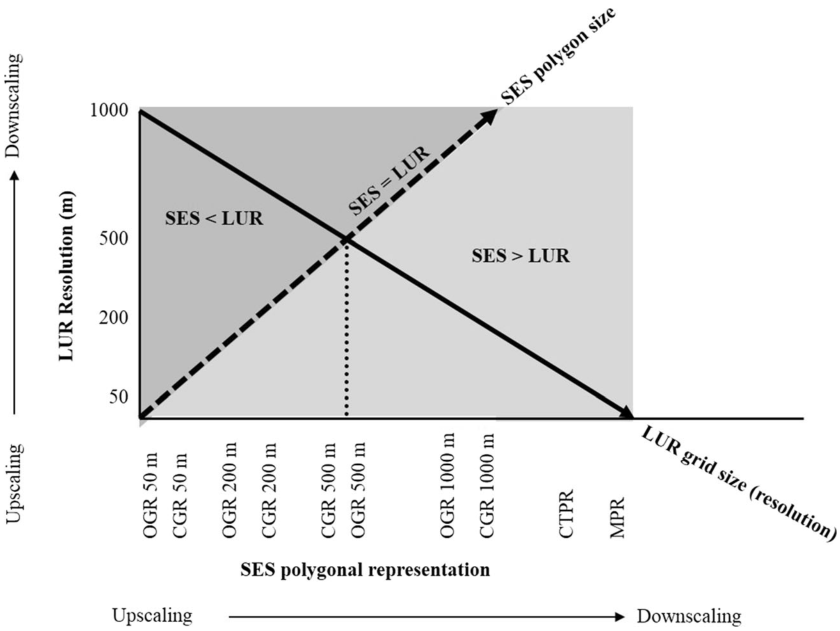

Fig. 1 Schematic spatial scales of the different aggregation methods used for SES (polygonal) and NO_x concentration (LUR grid) representations. The four polygonal SES representations are (a) municipality borders (*MPR*), (b) census tracts (*CTPR*), (c) ordered grid (*OGR*), and (d) clusters of grids with identical SES rank (*CGR*)

Air pollution

A land use regression (LUR) model (Levy et al., 2015) was used for calculating the annual average nitrogen oxide concentrations (NO_x) for the year of 2008. The model was developed for the whole Israel, using 104 monitoring stations of the national air quality monitoring network (AQMN). Briefly, the model examines explanatory variables including the national road network, traffic counts, population density, land use characteristics (residential and industrial), and vegetation cover (based on the normalized vegetation difference index, NDVI, at 30 m grid cell resolution). In addition, geolocation data (latitude, longitude, and elevation above the street level) of the monitoring stations were also used. A multivariate generalized additive model (GAM; implemented using the *mgcv* package in R version 2.15.1) was applied in a supervised forward stepwise regression in order to predict NO_x levels. Independent model variables include the road network, population density, elevation above ground, year of measurement, and type of monitoring site (near-road or general). The overall R^2 of the GAM model was 0.74, representing good performance. The LUR model was applied on a 50-m grid and downscaled to obtain three coarser grid resolutions of 200, 500, and 1000 m. The downscaling procedure was based on the majority value of NO_x concentrations at the 50 m resolution in the larger (coarser) grid cell, thus avoiding few local extreme values that were obtained since the LUR model was regressed also against data from transportation air quality monitoring stations. Data from such stations are known to reflect local conditions that are expected to spatially decay rather fast with the distance from the road. Using the mode rather than the mean circumvent the effect of these local high concentrations is at the coarser grid. This procedure enabled us to study relationships between data layers that do not stem from using distinct resolution (grid)-specific models and model inputs but result only from the varying spatial resolution of the data layers. Namely, the 50-m grid resolution LUR model estimates were used to obtain air quality estimates at any of the coarser resolutions. This data-aggregation approach enabled us to examine the effect of air quality estimates at different spatial resolutions while controlling for the effects that could arise in case different models are used for producing the estimates at the distinct spatial scales. It is noteworthy that other downscaling approaches require using aggregated model inputs and new parameterization of the LUR model (Beelen et al., 2013); thus, spatial downscaling and model downscaling will be mixed together.

Socioeconomic status

Three methods for mapping the SES polygonal boundaries were used in this study (Fig. 2). (A) Administrative boundaries systems—we used two alternative systems that are in use by the Israel CBS: (A1) the municipal polygonal representation (MPR), where the SES ranking is provided at 10 categories (1—low to 10—high). The boundaries of this polygonal system represent physical and man-made features, such as roads and railway lines (Maantay and Asthma and air pollution in the Bronx 2007; Stafford et al., 2008). (A2) The census tracts polygonal representation (CTPR), where the SES ranking is provided at 20 categories (1—low to 20—high). It is noteworthy that the definition of administrative polygonal representation (MPR, CTPR) is mostly identical across countries. In particular, census tracts are mostly as homogenous as possible in their land use, demographical attributes, and socioeconomic characteristics (Padilla et al., 2013).

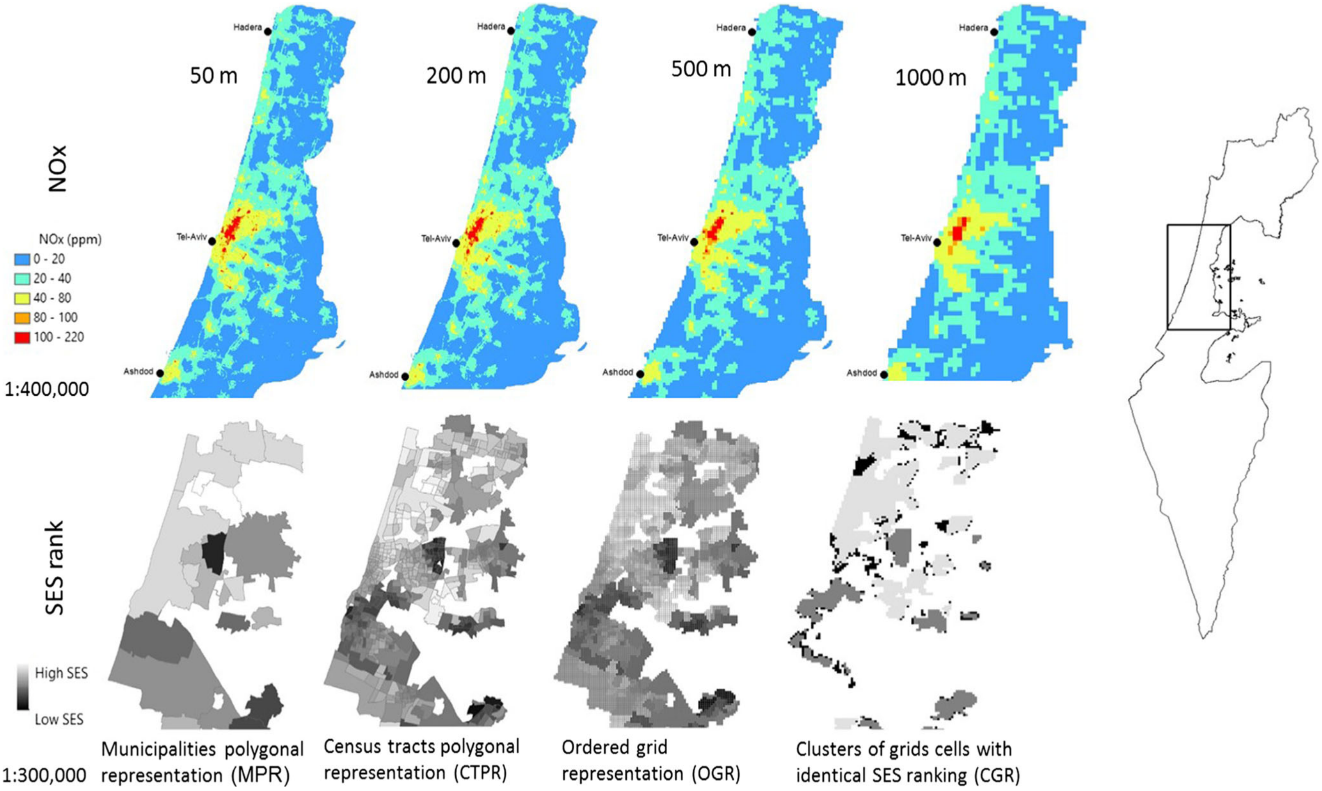
50 m
200 m
500 m
1000 m
Hadera
Tel-Aviv
Ashdod
NOx
NOx (ppm)
0 - 20
20 - 40
40 - 80
80 - 100
100 - 220
1:400,000
SES rank
High SES
Low SES
1:300,000
Municipalities polygonal representation (MPR)
Census tracts polygonal representation (CTPR)
Ordered grid representation (OGR)
Clusters of grids cells with identical SES ranking (CGR)

Fig. 2 Aggregation of spatial data. *Top*: LUR model results at different spatial grid resolutions (50, 200, 500, and 1000 m). The spatial heterogeneity decreases with the increase in the size of the grid cells. *Bottom*: different polygonal representations of the underlying population SES ranking based on the 2008 Israeli census (details of each representation are provided in the text). *Right*: study area overlaid on a schematic map of Israel

In general, the average population per census tract may vary between countries, in line with the total population, but due to its lower aggregation the CTPR averages information about the population to a lesser degree than the MPR. For instance, in Israel, census tracts represent areas with 3000–6000 inhabitants (Israel CBS) whereas in Canada and in the USA they consist of 2500–8000 inhabitants (Parenteau and Sawada, 2011) and 1200–8000 inhabitants (Bell and Ebisu, 2012), respectively. Furthermore, like the MPR boundaries, the borders of the CTPR also usually follow physical features. Consequently, the census tracts are not equal in their area or shape, which may lead to errors when linking layers of distinct spatial properties (Apparicio et al., 2008).

(B) Ordered grid representation (OGR)—we applied this method to obtain polygonal SES representation that match the shape and the size of the grid cells of the LUR model. Namely, this approach forces the size and shape of the grid to be consistent for the two variables (NO_x, SES) along the generalization procedure (i.e., the increase/decrease in the grid size). However, whereas the NO_x values were directly calculated by resampling the LUR model at the required resolution, the SES values were assigned based on the underlying census tract that covered the largest area fraction of the grid cell, using *zonal statistics* in ArcGIS 10.1.

(C) Clusters of grids cells with identical SES ranking (CGR)—to delineate homogenous polygons in terms of their SES, we used a semi-automated approach. First, based on the OGR approach, we obtained SES at four different grid cell sizes (50, 200, 500, and 1000 m) that matched the LUR model output, as described above. Next, a semi-automatic clustering process was used for grouping grid cells together using the Getis-Ord Gi* cluster analysis (*hotspots* in ArcGIS 10.1; Parenteau and Sawada, 2011). Namely, the grid cells at the SES layer were combined into clusters according to their *z*-score and *p* value as follows: (1) grid cells with *p* value >0.1 were excluded due to lack of statistically significant evidence for a cluster, assuming they represent a random spatial pattern. (2) A cluster of low SES ranking was assigned when the *p* value was <0.05 and the *z*-score was < -1.96. (3) A cluster of mid SES ranking was assigned when the *p* value was between 0.05 and 0.1 and the *z*-score was between −1.96 and −1.65 or between 1.96 and 1.65, and (4) a cluster of high SES ranking was assigned when the *p* value was <0.05 and the *z*-score was >1.96. Inherently, this algorithm gives more weight to the tails of the SES distribution (i.e., the very high or the very low SES). It is noteworthy that after the clusters were formed via aggregation, their polygonal shape and size vary. It is noteworthy that the aggregation of SES patterns, like the aggregation of the NOx patterns, did not require new data but rather involved warping the existing data in different ways that result in spatial aggregation.

Linking data layers

The level of ecological bias was examined in response to the varying spatial aggregation of the underlying NO_x and SES polygonal representations. The relationships between the different data layers were assessed as follows. First, we defined the three polygonal SES representations as described above. Next, the socioeconomic ranking of the underlying population was assigned to each representation based on the 2008 Israeli census. Finally, NO_x concentrations were assigned to each SES polygonal representation according to the rules detailed below, which reflect the spatial relationships between the LUR model grid size and the SES polygonal representation. It is important to note that the last two steps involved sometimes further downscaling, to obtain a "joint" spatial resolution of the two variable layers. Specifically, the latter step was performed only when different data layer structures were examined, representing real cases of distinct sampling strategy, model output, or data aggregation as a result of, e.g., privacy concerns. Figure 1 summarizes the different data representations of the exposure metric (LUR estimates) and the confounder variable (SES polygonal representation), illustrating schematically how the change in SES polygonal and LUR (NO_x) grid cell sizes can lead to downscaling of the spatial information. When the size and shape of the SES polygons and the LUR grid cell coincide, each SES polygon was assigned the NO_x concentration of the LUR grid cell that is exactly underneath it. However, when the LUR grid cells are larger than the SES polygons (dark gray area in Fig. 1) each grid cell with a LUR estimated NO_x concentration is assigned the weighted average SES ranking of the SES polygons that lie under it. On the other hand, when the SES polygons are larger than the LUR grid cells (light gray area in Fig. 1) each SES polygon is assigned the weighted average NO_x concentration of the LUR grid cells that lie under it.

Statistical methods

To examine to what extent the spatial representation of the two data layers (risk factor and confounder) affects the statistical relationship between them, we calculated the Spearman correlation coefficient (r) for all the combinations of NO_x and SES spatial generalizations. For simplicity, the SES ranking was divided into three categories: high, medium, and low. When using the MPR, CTPR, and OGR representations, the SES categories were divided according to the natural breaks method, which provides optimal grouping of similar values while maximizing the differences between groups (ArcGIS 10.1). For the CGR representation, SES categories were divided into three classes according to their z-scores and p-values, which were obtained as part of the cluster analysis test (see above).

The confounding relationships between SES and NO_x across the whole study area were studied using logistic regression and one-way ANOVA, to determine if spatial variation of NO_x concentrations among different SES groups and polygonal representations is evident. High F-statistics indicate high spatial variability of exposure to NO_x among the SES groups, which reduces the potential for ecological bias. Differences among SES groups were further examined using Duncan's multiple range post hoc test (Duncan 1995; STATISTICA 8 package with the General Linear Model (GLM) procedure, using the Mean option).

Results

Figure 3 depicts the Spearman correlation coefficient of all the SES–NO_x representation pairs. The correlations between NO_x concentrations and the MPR or CTPR SES representations were insignificant (p value >0.05). In contrast, significant (p value <0.05) although low correlations were obtained between the OGR and CGR SES representations and NO_x despite the fact that these representations involved up/downscaling of the SES ward structure, as opposed to the MPR and CTPR representations. The relatively small grid size, especially of the OGR polygonal SES representation, required a large number of grid cells to map the study area and, therefore, implicates a small number of degrees of freedom. This may decrease the robustness of the results, since a low signal-to-noise ratio (SNR) may mask the NO_x–SES interaction due to artificially increased variability. Small unit area may also increase the probability of spatial autocorrelation. Namely, as the number of polygons increases their size decreases, and pairs of polygons which are adjacent to each other are more probable to show similar values (Dormann et al., 2007). It is noteworthy that in this case the autocorrelation does not represent a real spatial pattern of the SES groups but rather only the experimental design. Indeed, Fig. 3 reveals that changes in the spatial resolution of both the LUR model and the SES polygonal representation alter the statistical correlation.

The Spearman correlation coefficient (r) increased with the increase in size of the grid cells that underlie the CGR and OGR SES representations. For example, r increased from 0.03 to 0.09 for the OGR SES representation when LUR_{50} was used for exposure estimation (Fig. 3). Likewise, r increased dramatically from 0.04 to 0.26 for the CGR SES representation when LUR_{50} was used for exposure estimation. These results demonstrate possible ecological bias that could result when using data layers with different spatial resolutions as a result of distinct re-scaling procedures (down/upscaling) or aggregation schemes.

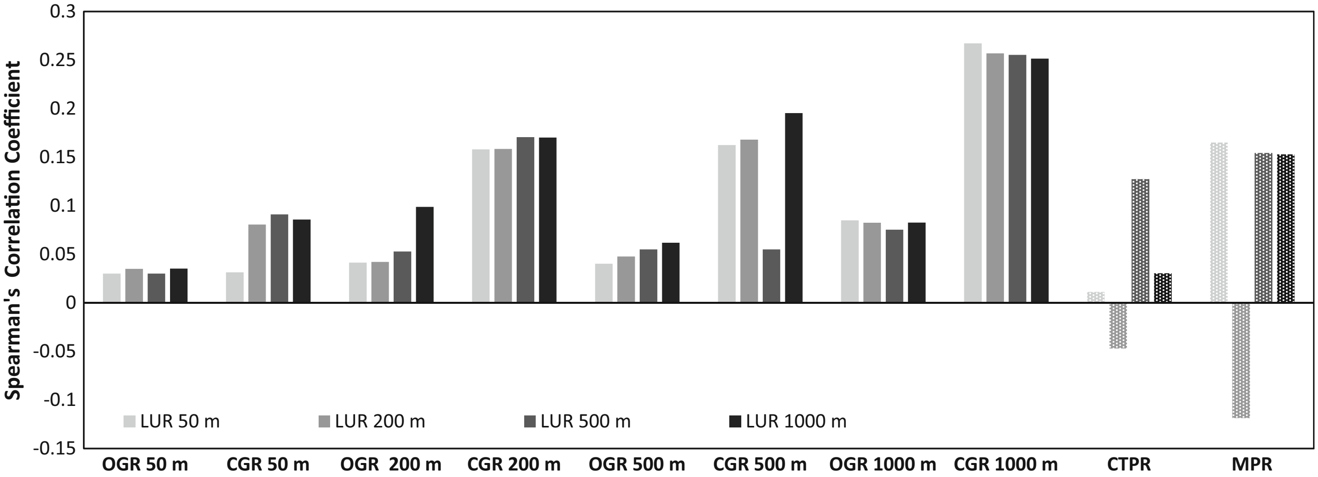

Fig. 3 Effect of the polygonal representation and of the LUR grid resolution on Spearman's correlation between the underlying population SES ranking (the confounder) and the NO_x concentrations (the risk factor). *Solid fill*—significant correlation coefficients (p value <0.01), *dotted fill* (CTPR and MPR)—non-statistically significant correlations

An alternative scheme for studying the effect of spatial resolution on inter data-layer relationships is by ordinal logistic regressions between any of the commonly used SES categories (low, medium, high) of the different SES representations and the spatial resolution of the LUR-based exposure metric. Figure 4 shows the results of such logistic regression models, with SES being the categorical dependent variable and NO_x concentrations the continuous predictor. The results imply that in some cases significant associations between the NO_x levels and the SES groups were obtained (OR >1, where OR is the odds ratio), in particular when the polygonal representations (mainly OGR) involved up/downscaling procedures, and when the spatial resolution of the SES data was high. However, in general no clear pattern or a particular advantage was revealed for any of the polygonal SES representations. It should be further noted that using the administrative polygonal representations did not reveal any significant relationships with the ambient NO_x levels.

One possible reason for the observed decrease in the strength of the associations is the limited spatial variability of NO_x concentrations, which may result from the LUR model grid resolution. Namely, when there is no spatial variability of the NO_x concentrations no differential exposure is expected at the individual-level analysis and no interaction could be expected between any of the SES representations and the NO_X concentrations. To explore this, we ran a one-way ANOVA and determined the extent of the spatial variability of NO_x concentrations for different SES groups (low, medium, high) and SES representations (Table 1). A low F-statistics implies low variability of the NO_x concentrations among the different SES polygonal representations, indicating a potential for effect measure errors due to aggregation of spatial data. Table 1 reveals no spatial variability of the NO_x concentrations when the traditional SES polygonal representation (MPR and CTPR) were used. In contrast, when the spatial resolution of both the LUR and the SES maps was high (i.e., small grid cells), the spatial variability (heterogeneity) of the NO_x concentrations was large among the SES categories (e.g., a SES grid size of 50 m and a LUR grid between 50 and 500 m, Table 1). Moreover, Table 1 reveals large a variation in the NO_x concentrations among the CGR SES representations, which decreases from ~700 to about zero with the increase in the SES grid size from 50 to 1000 m. This represents the common notion that the spatial resolution of the risk factor (here NO_x concentration estimates) may significantly bias the associations in environmental epidemiology studies.

Discussion

Usually, epidemiological studies on air pollution and its related health effects do no account for uncertainties in the estimated exposure (Sheppard et al., 2012). As shown in this work, some of this uncertainty may be attributed to spatial aggregation of the environmental risk factors and the confounding parameters (Jerrett et al., 2010; Greenland and Morgenstern, 1989). Since nowadays, environmental risks are mostly investigated using individual-level information (Idrovo, 2011), the aggregation process may lead to an ecological bias. To date, ecological bias in epidemiological studies has been mostly related to the ignored variability among individuals that consist a group for which identical exposure and confounders are assigned. For example, individuals that reside within a given area (e.g., census tract) that are assigned the same exposure or risk factors. Similarly, individuals that belong to the same SES group are assumed to be affected to the same degree by an identical confounder. However, in practice, the exposure of these individuals may be highly variable (Wakefield, 2008). This study demonstrates how different aggregation schemes of the underlying risk factors and

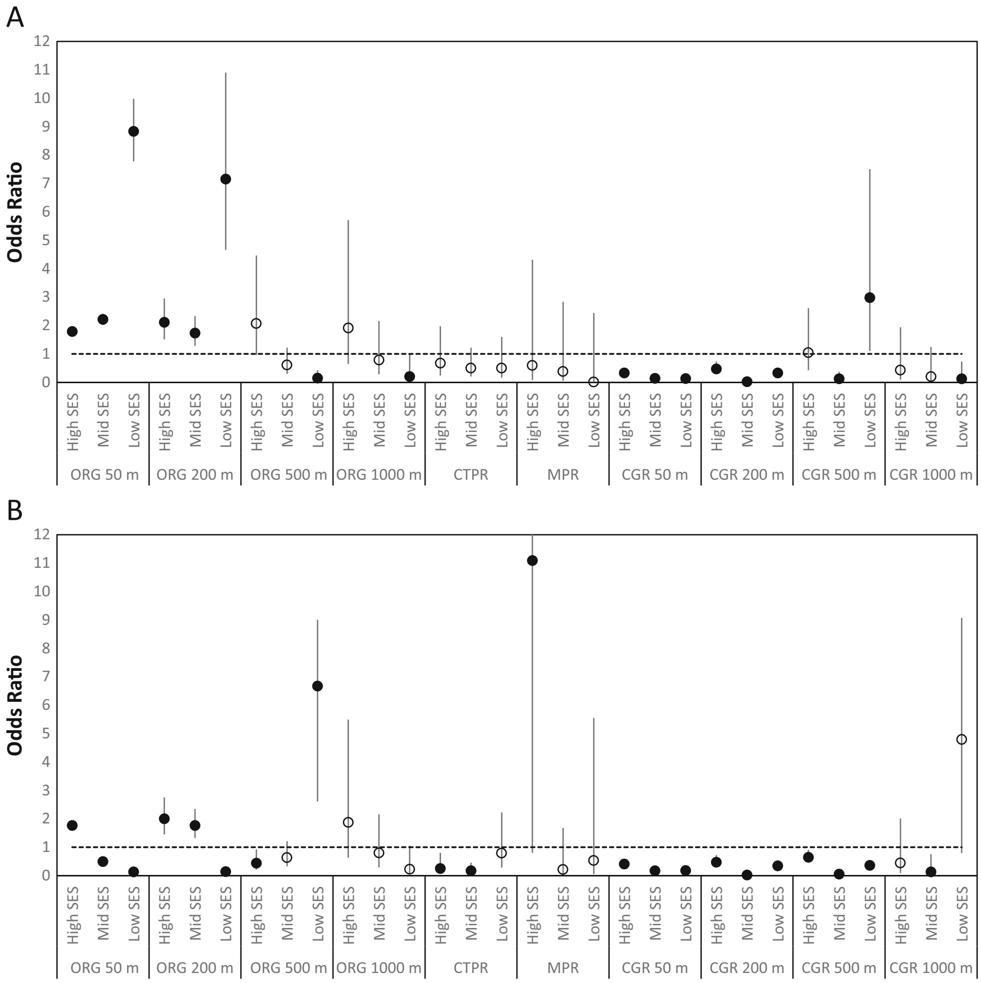

Fig. 4 Odd ratios of exposure metrics (NO_x concentrations) obtained from a LUR model at different grid sizes (**a** 50 m, **b** 200 m, **c** 500 m, and **d** 1000 m) and socioeconomic groups (high SES, medium SES, and low SES) based on different SES polygonal presentations (municipality borders (MPR), census tracts (CTPR), ordered grid (OGR), and equal SES clusters (CGR), see the text for further description. *Filled symbols* are statistically significant; *open symbols* are statistically non-significant

confounders can modulate the associations and statistical inference (both its magnitude and direction) between exposure metrics and confounders within an environmental health study. In particular, we show that merging of environmental information that is obtained at different spatial resolutions and polygonal representation can lead to erroneous results in terms of either over-prediction or under-prediction relative risks.

Furthermore, this study explores various mechanisms by which up/downscaling of spatial information sources can introduce bias. Specifically, we showed that up/downscaling may reduce the variability of the exposure metric (here, NO_x concentrations) among equal confounder-level groups (here, the SES ranking) and eliminates (or drastically reduces) the statistical association between them (Fig. 3). For example, we found that in most cases the traditional wards for which SES is available, CTPR and MPR, involve downscaling of the environmental information, which leads to non-significant statistical associations with the exposure metrics. For example, the extent of the polygonal CTPR and MPR representation of the SES leads to averaging of the NO_x concentrations over large areas and results in low exposure variability among the SES groups.

The question to what extent the choice of specific polygonal boundaries affects the averaging and uniformity of the exposure estimates is case specific. While some studies reported bias due to specific choices of the polygons' boundaries (Flowerdew et al., 2008; Poortinga et al., 2008), other found no such influence across different polygonal representations (Stafford et al., 2008; Lovasi et al., 2008; Eitan et al., 2010). To avoid the problem of inconsistent shape and size of the polygons' borders, some studies suggested that the finest polygonal unit should be used for the analysis, unless prior evidence indicates that larger units are essential for investigating the research question (Jerrett et al., 2010). Thus, naively, one could expect that the highest spatial variability of both the SES ranking and the NO_x concentrations will be obtained when using the finest spatial resolution. If generally true, this should have been revealed when the LUR model results (NO_x) and

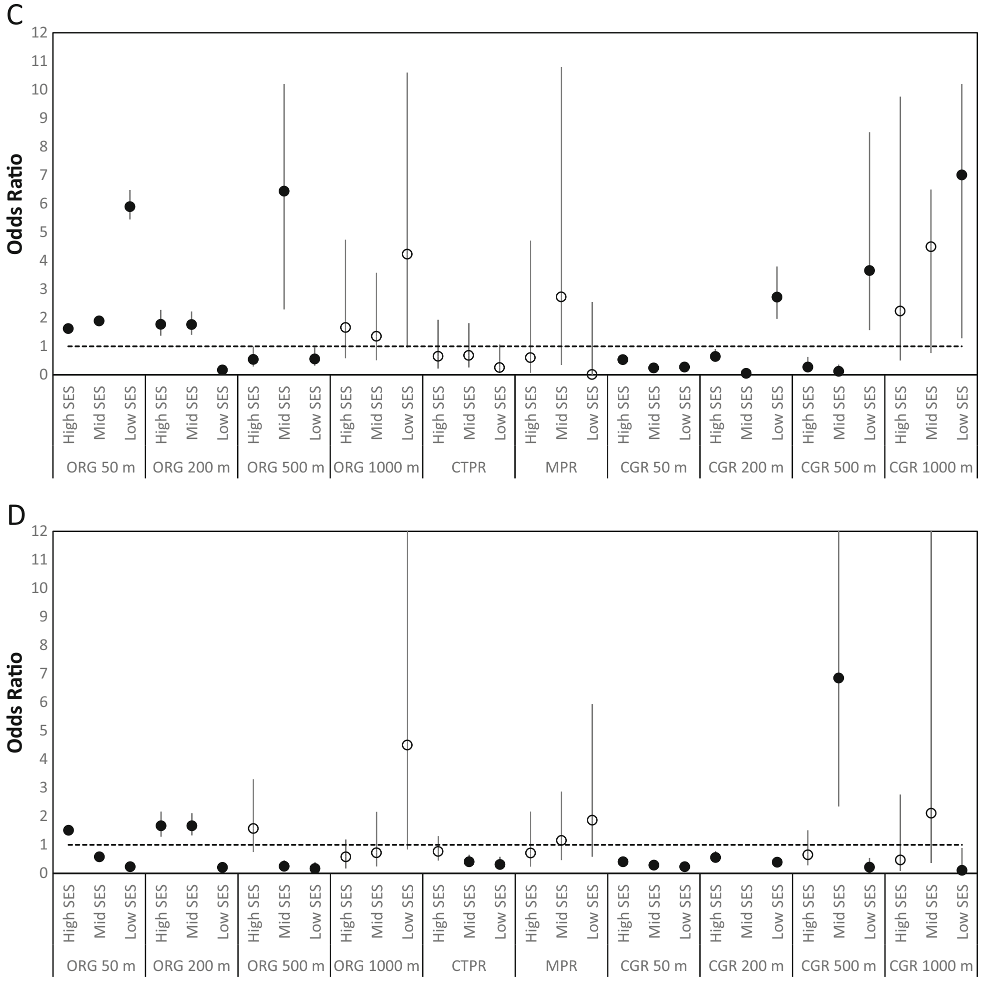

Fig. 4 (continued)

the OGR or CGR polygons (SES) at the 50 m grid spatial resolution were used. However, our results show that the highest spatial heterogeneity over the whole study area was obtained when the NO_x concentrations were presented on a 500-m grid and the SES OGR categories were presented on a 50-m grid (Table 1). Hardly any spatial heterogeneity of the exposure (in terms of NO_x concentrations) was seen when using the traditional SES polygonal representations (MPR, CTPR). Hence, the spatial variability of both the risk factors and the confounders is clearly affected by the subjective choices of the user regarding their spatial representation. Area-level confounders and exposures, as any other spatial

Table 1 F-statistic values of the one-way ANOVA test between the SES polygonal representation and the LUR grid resolution. Higher F-statistics represent higher spatial variation of the NO_x concentrations among the SES categories

Polygonal SES representation	LUR 50 m	LUR 200 m	LUR 500 m	LUR 1000 m
CTPR	2.56	2.11	1.95*	11.14
MPR	0.70	1.15	0.62	0.56
OGR 50 m	590.6*	1559.0*	2576.3*	6.14*
OGR 200 m	101.9*	103.2*	103.81*	6.35*
OGR 500 m	16.67*	16.68*	17.17*	6.44*
OGR 1000 m	1467.83	93.03*	16.16	5.28*
CGR 50 m	716.5*	695.1*	672.0*	717.4
CGR 200 m	95.5*	99.87*	105.23*	105.0*
CGR 500 m	7.67*	8.36*	9.69*	41.0*
CGR 1000 m	3.36*	3.25*	3.14*	2.85*

Statistically significant values are marked by * phenomena, have characteristic scales and resolutions (Fekete et al., 2010). In this work, we show that user decisions regarding the polygonal/grid representation of the various data layers should account for the spatial distribution of the phenomena studied. These decisions can substantially affect the statistical relationships between the model variables (stressors and confounders) and modify the correlation with the dependent variable, as well as their specific contributions to its explained variability.

As clearly demonstrated in this work and as known from the literature, there is always an unavoidable tension between two opposing strategies: data aggregation (to enhance the statistical significance) and generalization of ecological information (Jerrett et al., 2010). Since spatially aggregated information displays a higher level of uncertainty than individual-level data, observed patterns may contain artifacts of the aggregation process (Greenland and Morgenstern, 1989). For example, in Tel Aviv (the largest city in our study area), Israel, none of the SES polygonal representations studied in this work led to a consistent confounding association with the NO_x concentrations. This result is in contradiction with previous findings, which suggested that in the Tel Aviv area there is a geographic overlap between areas that are characterized by higher NO_x concentrations and the higher SES ranking statistical wards (Myers et al. 2013). In general, both epidemiological and environmental data are available at relatively crude spatial resolutions, and their aggregation may lead to considerable generalization of the individual-level risks and to ecological bias. We demonstrated here that one source of this ecological bias can be easily quantified, as opposed to ecological bias that results from unaccounted for individual-level risk factors and confounders (Sheppard et al., 2012; Goodman et al., 2011). In general, our results suggest that using broader scales that generalize spatial information may alter (mostly weaken) the relationships between environmental stressors and health outcomes. It should be noted that while our findings may be context-specific, they demonstrate an important universal principle, namely that the spatial representation of the geocoded data has an impact on the derived relationships between different layers of spatial information.

Acknowledgments This research was supported by the Technion Center of Excellence in Environmental Sciences and Environmental Health (TCEEH). We thank Yuval, Y. Etzion, and B. Fishbain for fruitful discussions.

References

Amster ED, Haim M, Dubnov J, Broday DM (2014) Contribution of nitrogen oxide and sulfur dioxide exposure from power plant emissions on respiratory symptom and disease prevalence. Environ Pollut 186:20–28

Apparicio P, Abdelmajid M, Riva M, Shearmur R (2008) Comparing alternative approaches to measuring the geographical accessibility of urban health services: distance types and aggregation-error issues. Int J Health Geogr 7:7

Beelen R, Hoek G, Vienneau D, Eeftens M, Dimakopoulou K, Pedeli X, Eriksen KT (2013) Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe—the ESCAPE project. Atmos Environ 72:10–23

Bell ML, Ebisu K (2012) Environmental inequality in exposures to airborne particulate matter components in the United States. Environ Health Perspect 120:1699–1704

Cyril S, Oldroyd JC, Renzaho A (2013) Urbanisation, urbanicity, and health: a systematic review of the reliability and validity of urbanicity scales. BMC Public Health 13:513

Diez Roux AV, Mair C (2010) Neighborhoods and health. Ann N Y Acad Sci 1186:125–145

Dormann FC, McPherson MJ, Araújo BM, Bivand R, Bolliger J, Carl G, Davies GR, Hirzel A, Jetz W, Daniel Kissling W, Kühn I, Ohlemüller R, Peres-Neto RP, Reineking B, Schröder B, Schurr MF, Wilson R (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30: 609–628

Duncan D (1995) Multiple range and multiple F tests. Biometrics 11:1–42

Eitan O, Yuval, Barchana M, Dubnov J, Linn S, Carmel Y, Broday DM (2010) Spatial analysis of air pollution and cancer incidence rates in Haifa Bay, Israel. Sci Total Environ 408:4429–4439

Fekete A, Damm M, Birkmann J (2010) Scales as a challenge for vulnerability assessment. Nat Hazards 55:729–747

Flowerdew R, Manley DJ, Sabel CE (2008) Neighborhood effects on health: does it matter where you draw the boundaries? Soc Sci Med 66:1241–1255

Goodman A, Wilkinson P, Stafford M, Tonne C (2011) Characterising socio-economic inequalities in exposure to air pollution: a comparison of socio-economic markers and scales of measurement. Health and Place 17:767–774

Greenland S, Morgenstern H (1989) Ecological bias, confounding, and effect modification. Int J Epidemiol 18:269–274, Erratum in Int J Epidemiol 1991 20(3):824

Gryparis A, Paciorek CJ, Zeka A, Schwartz J, Coull BA (2009) Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics 10:258–274

Hoskins AJ, Bush A, Gilmore J, Harwood T, Hudson LN, Ware C, Williams KJ, Ferrier S (2016) Downscaling land-use data to provide global estimates of five land-use classes. Ecology Evolution 6: 3040–3055

Idrovo AJ (2011) Three criteria for ecological fallacy. Environ Health Perspect 119:A332

Jerrett M, Gale S, Kontgis C (2010) Spatial modeling in environmental and public health research. Int J Environ Res Public Health 7:1302–1329

Levy I, Levin N, Yuval, Schwartz JD, Kark JD (2015) Back-extrapolating a land use regression model for estimating past exposures to traffic-related air pollution. Environ Sci Technol 49:3603–3610

Lovasi GS, Moudon AV, Smith NL, Lumley T, Larson EB, Sohn DW, Siscovick DS, Psaty BM (2008) Evaluating options for measurement of neighborhood socioeconomic context: evidence from a myocardial infarction case–control study. Health Place 14:453–467

Maantay J (2007) Asthma and air pollution in the Bronx: methodological and data considerations in using GIS for environmental justice and health research. Health Place 13:32–56

Myers V, Broday DM, Steinberg DM, Yuval, Drory Y, Gerber Y (2013) Exposure to particulate air pollution and long-term incidence of frailty after myocardial infarction. Ann Epidemiol 23:395–400

Openshaw S (1994) The modifiable area unit problem. Concepts Tech Mod Geogr 38:1–41

Padilla CM, Deguen S, Lalloue B, Blanchard O, Beaugard C, Troude F, Navier DZ, Vieira VM (2013) Cluster analysis of social and environment inequalities of infant mortality. A spatial study in small areas revealed by local disease mapping in France. Sci Total Environ 454–455:433–441

Parenteau M-P, Sawada MC (2011) The modifiable areal unit problem (MAUP) in the relationship between exposure to NO2 and respiratory health. Int J Health Geogr 10:58

Poortinga W, Dunstan FD, Fone DL (2008) Neighbourhood deprivation and self-rated health: the role of perceptions of the neighbourhood and of housing problems. Health Place 14:562–575

Ryan PH, LeMasters GK, Levin L, Burkle J, Biswas P, Hu S, Grinshpun S, Reponen T (2008) A land-use regression model for estimating microenvironmental diesel exposure given multiple addresses from birth through childhood. Sci Total Environ 404:139–147

Shaddick G, Lee D, Wakefield J (2013) Ecological bias in studies of the short-term effects of air pollution on health. Int J Appl Earth Obs Geoinf 22:65–74

Sheppard L, Burnett RT, Szpiro AA, Kim S-Y, Jerrett M, Pope CA, Brunekreef B (2012) Confounding and exposure measurement error in air pollution epidemiology. Air Qual Atmosphere Health 5:203–216

Stafford M, Duke-Williams O, Shelton N (2008) Small area inequalities in health: are we underestimating them? Soc Sci Med 67:891–899

Wakefield J (2008) Ecologic Studies Revisited. Annu Rev Public Health 29:75–90

Wakefield J, Shaddick G (2006) Health-exposure modeling and the ecological fallacy. Biostatistics 7:438–455

Yuval, Broday DM (2006) High-resolution spatial patterns of long-term mean concentrations of air pollutants in Haifa Bay area. Atmos Environ 40(20):3653–3664.

Zou B, Wilson JG, Zhan FB, Zeng Y (2009) Air pollution exposure assessment methods utilized in epidemiological studies. J Environ Monit 11:475