Confidential: Do Not Cite or Distribute

Status Report:
Land Use Regression in New York City

May 27, 2005

Prepared by:
Zev Ross, MS, ZevRoss Spatial Analysis
Michael Jerrett, PhD, University of Southern California

ZevRoss
Spatial Analysis
904 Giles Street, Ithaca, NY 14850
(607) 277 0004 phone
(866) 877 3690 fax, toll free
info@zevross.com • www.zevross.com

Background

Several recent studies have demonstrated that regression models of air pollutants using traffic and land use as independent variables can predict air pollutant levels at least as well, and in some cases significantly better, than more complicated dispersion modeling. This is particularly true for more local pollutants such as nitrogen dioxide, but is also true for pollutants with regional patterns such as particulate matter.

In European studies for example, researchers were able to use traffic and land use nearby monitors to explain 79-87 percent of the variation in nitrogen dioxide and 50-73 percent of the variation in PM2.5 (Brauer*, et al.* 2003, Briggs*, et al.* 2000). In San Diego, California we were able to construct a model based on the methods developed by Briggs *et al* that explained 79% of the variation and predicted 12 validation samples to within, on average, 2.1 ppb (Ross*, et al.* In Review).

Land use regression, as the method has come to be known, uses GIS to calculate the traffic, land use, demographics and topography within a given distance (or several distances) of an air pollutant monitoring station. These buffer variables, in turn, are used in a multivariate regression and the resulting equation is used to predict air pollutant levels at unknown locations.

New York City Metropolitan Area Land Use Regression

In the New York City (NYC) metropolitan area we have made significant progress in developing a land use regression model for fine particulate matter. The equations themselves have yet to be developed, but the data collection and GIS calculations have been nearly completed.

Particulate Matter Data

For the NYC model, we have calculated three-year averages (1999-2001) for all of the monitors in a 28 county area of NYC using daily fine particulate matter data from the EPA's Air Quality Subsystem. We first computed quarterly averages for each year for monitors that had at least 8 observations (i.e., at least half of scheduled monitoring days). We then computed three-year averages by taking an average of the quarterly means when a complete set of four quarterly means existed. In total we have averages for 62 monitors in 22 counties (6 counties have inadequate monitoring data). Thirty-five of these monitors are located in New York, seventeen in New Jersey and ten in Connecticut.

The PM25 data for these 62 sites is normally distributed and ranges from a low of 11.10 to a maximum of 19.60 with a mean of 14.31 $\mu g/m^3$. High values were generally centered on Manhattan and eastern New Jersey counties with concentrations decreasing with increasing distance from New York City (see map of inverse distance values).

Historical fine particulate matter estimates will be calculated using methods developed by Ramona Lall and others at NYU (Lall*, et al.* 2004) for the ten monitoring locations identified as having adequate PM2.5 and PM10 data.

Each of the 62 monitors has been buffered at 40, 100, 300, 500 and 1000 meters (see example in Map 1 and locations of all monitors in Map 2). For each of the buffers at each monitor we have calculated traffic and road density, land use at several resolutions, census data at the block group level and point source emissions.

Traffic Data

There are several sources of traffic data for New York City and surrounding counties, all have their advantages and disadvantages.

We initially considered using one of two sources of data. The New York State Department of Transportation (DOT) has traffic data available for major roads in New York State. This data, however, does not include adequate detail in New York City for an intra-urban analysis. As an alternative, the New York City DOT has significantly more detailed traffic data available. Unfortunately, while the traffic counts themselves are conducted on road *segments*, the data is represented in their database at nodes (intersections) rather than by segment. What further complicates use of this dataset is the fact that several traffic counts at several different segments may be represented at the same node. For example, if four counted segments meet at a node, all four counts will be identified with that node with no way of distinguishing which count goes with which segment.

Due to these limitations of state and city-level data we ultimately chose to use traffic estimates calculated by the New York Metropolitan Transportation Council (NYMTC). The NYMTC developed a transportation "Best Practices Model" to meet the federal requirements for long-range transportation planning for the entire metro area. The study area includes 28 counties in New York, New Jersey and Connecticut (the entire area of our analysis) and includes detailed traffic estimates for cars and trucks by segment. The 2002 model update was completed in early 2005. While, the NYMTC data does not provide estimates for all segments in New York City, New Jersey or Connecticut it includes enough detail to look at intra-urban traffic variation.

Land Use Data

We have identified three sources of land use data that will be considered in the land use regression analysis, these include an extremely detailed layer of taxlot data for the five boroughs of New York City, a 1:40,000 scale land cover layer for New Jersey and a 1:100,000 layer of land cover data that covers the entire region.

The taxlot data provided by the New York City Department of City Planning can be considered the gold standard of land use data. The 2004 data includes detailed GIS layers representing all 850,000 taxlots in the five counties and can be associated with tabular data on land use, zoning, assessed value and many other variables.

Unfortunately, similar fine resolution data on such a large scale does not exist for surrounding areas. New Jersey counties Hudson and Bergen, for example, are working on, but have not completed county-wide tax parcel maps with land use and/or zoning data. Westchester County similarly does not have county-wide parcel data (although individual municipalities do). And

Nassau County charges exorbitant amounts ($40,000 for parcel data countywide) for its data. The finest resolution land use data for New Jersey that we could identify was 1:40,000 data from the New Jersey Department of Environmental Protection available by watershed for the entire state.

Finally, we are also using the 1992 National Land Cover Data from the USGS. This data is relatively old and it is low resolution (1:100,000), but it covers the entire study area with a consistent set of land use categories. Given that regional sources make a significant contribution to particulate matter concentrations, high resolution representations of land use may not be necessary and the NLCD maps may be appropriate.

All of the land use layers above contain a significant number of land use categories. For the land use regression we are limiting the analysis to relevant categories such as “industrial” and “high density residential.” A preliminary analysis also suggests that reducing the number of land use categories using principal components analysis may be a valuable approach.

Census Data

We are also looking at neighborhood demographics for the areas around air monitors. We are using Census block group-level data from the US Census bureau including total population, number of households, number of housing units, average family income and others.

Emissions Data

In an effort to account for regional influences on particulate matter concentrations we are also including 1999 emissions data from EPA’s National Emission Inventory. We have assembled particulate emissions data at both the county-level and for point sources. As part of the analysis we will look at including overall county emissions as a predictor variable as well as the number of facilities and total particulate emissions from these facilities within given distances.

Status

As of today the data collection and processing is nearly complete and ready for modeling. There are two pieces we are still waiting for, however, both of which are expected soon. First, the traffic data from the NYMTC sent to us in January had several errors that needed to be corrected. On May 26, NYMTC sent us revised traffic files in TransCAD format. These files will need to be converted to shapefiles for analysis. This conversion is expected by the end of the first week in June.

Finally, we are working on verifying the locations of air monitors in EPA’s Air Quality Subsystem. The location information in this database is notoriously error-prone and often does not include the most up-to-date information. As a result, we have contacted staff in both the NY DEC, NJ DEP and CT DEP for updated location information. As of May 17, all three of these agencies have provided revised location data. In all cases, these new locations were determined by either GPS or aerial photos. In New York, they updated locations with both GPS and laptops equipped with Arcview in the field.

References

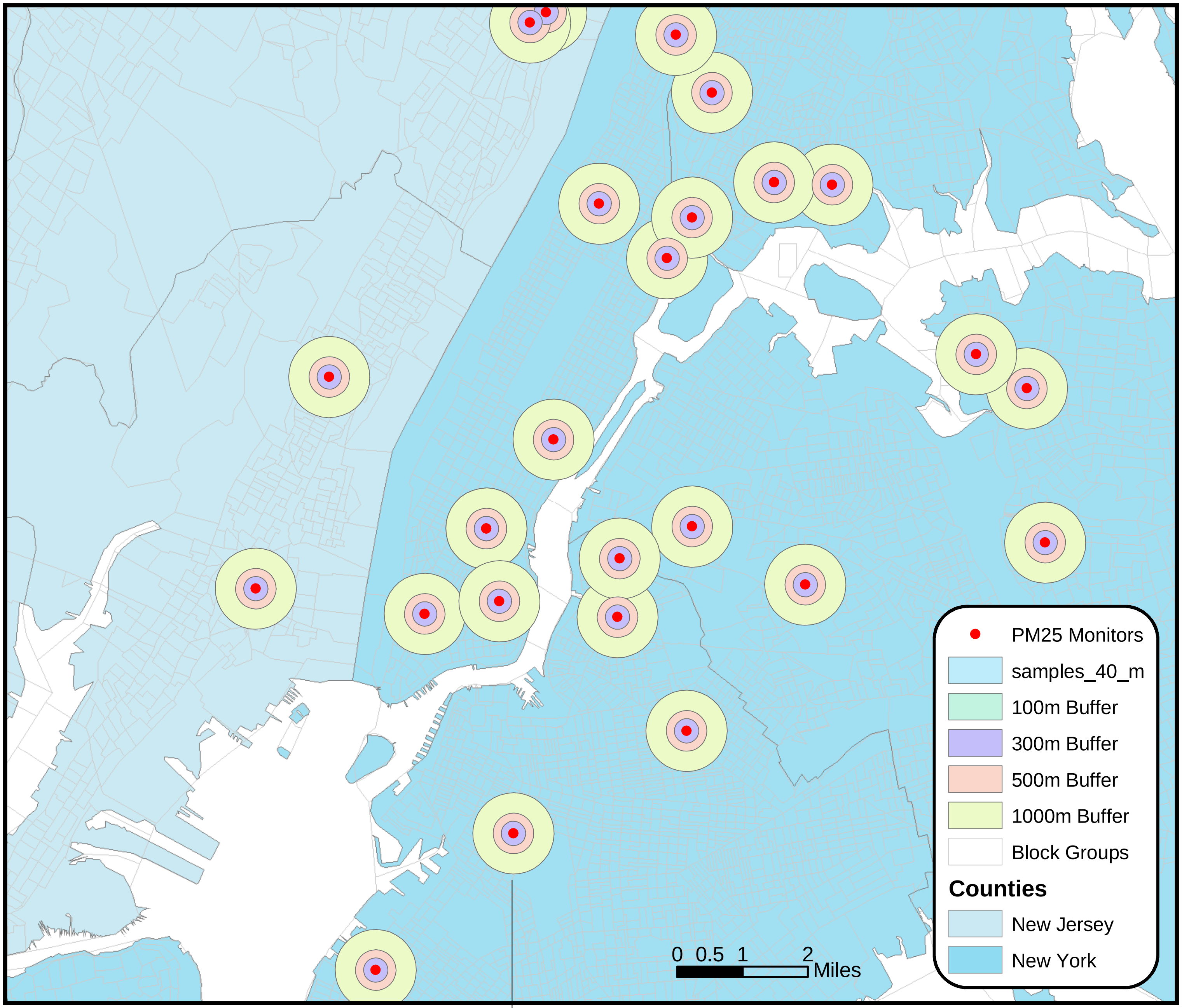
1. Brauer M, Hoek G, van Vliet P, Meliefste K, Fischer P, Gehring U, Heinrich J, Cyrys J, Bellander T, Lewne M, Brunekreef B. 2003. Estimating long-term average particulate air pollution concentrations: Application of traffic indicators and geographic information systems. Epidemiology 14:228-239.
2. Briggs DJ, de Hoogh C, Guiliver J, Wills J, Elliott P, Kingham S, Smallbone K. 2000. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. Science of the Total Environment 253:151-167.
3. Lall R, Kendall M, Ito K, Thurston GD. 2004. Estimation of historical annual $PM_{2.5}$ exposures for health effects assessment. Atmospheric Environment 38:5217-5226.
4. Ross Z, English P, Jerrett M, Scalf R, Gunier RB, Smorodinsky S, Wall S. In Review. Nitrogen dioxide prediction in southern California using land use regression modeling: potential for environmental health analyses. Journal of Exposure Analysis and Environmental Epidemiology.

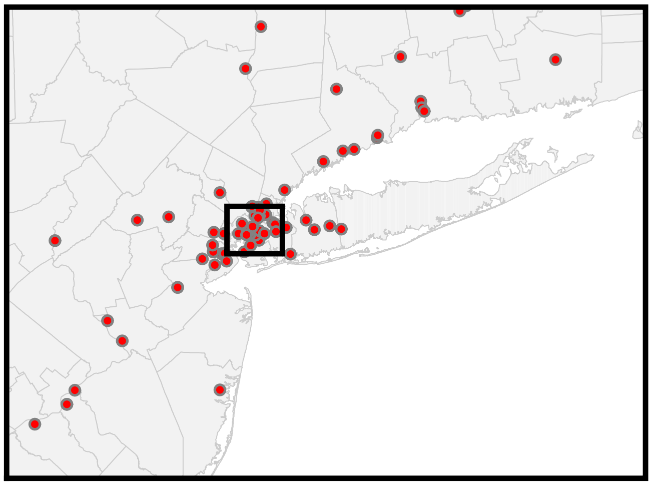
Map Descriptions:

1. A simple map that illustrates the buffering and clipping of GIS layers within a given distance of PM2.5 monitors.
2. The locations of PM2.5 monitors in the New York City metropolitan area.
3. Fine particulate matter sample values and inverse distance weighting.
4. The extent of the traffic layer from NYMTC in Manhattan. While the example shown is based on 1996 data, the model has been updated to include 2002 data. This data is available for all 28 counties in the region and is available for four time periods (AM, PM, Midday and Night).
5. Example of tax lot data for New York City.
6. One of the land use layers. This older National Land Cover Data layer is relatively low resolution but covers the entire research area. We are also evaluating the value of using NYC tax lot data which has a land use attribute and is much, much finer resolution.

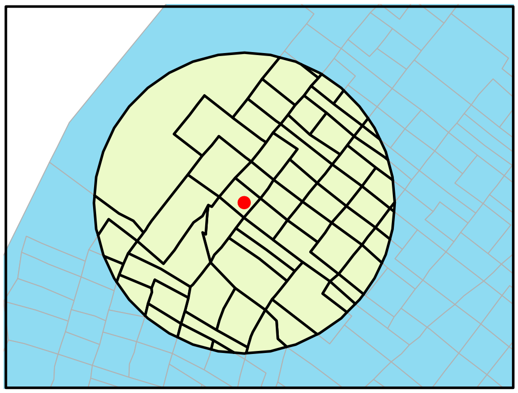
An Example of Clipping and Buffering for Land Use Regression:

Particulate Matter Monitoring Stations and Census Block Groups in New York City





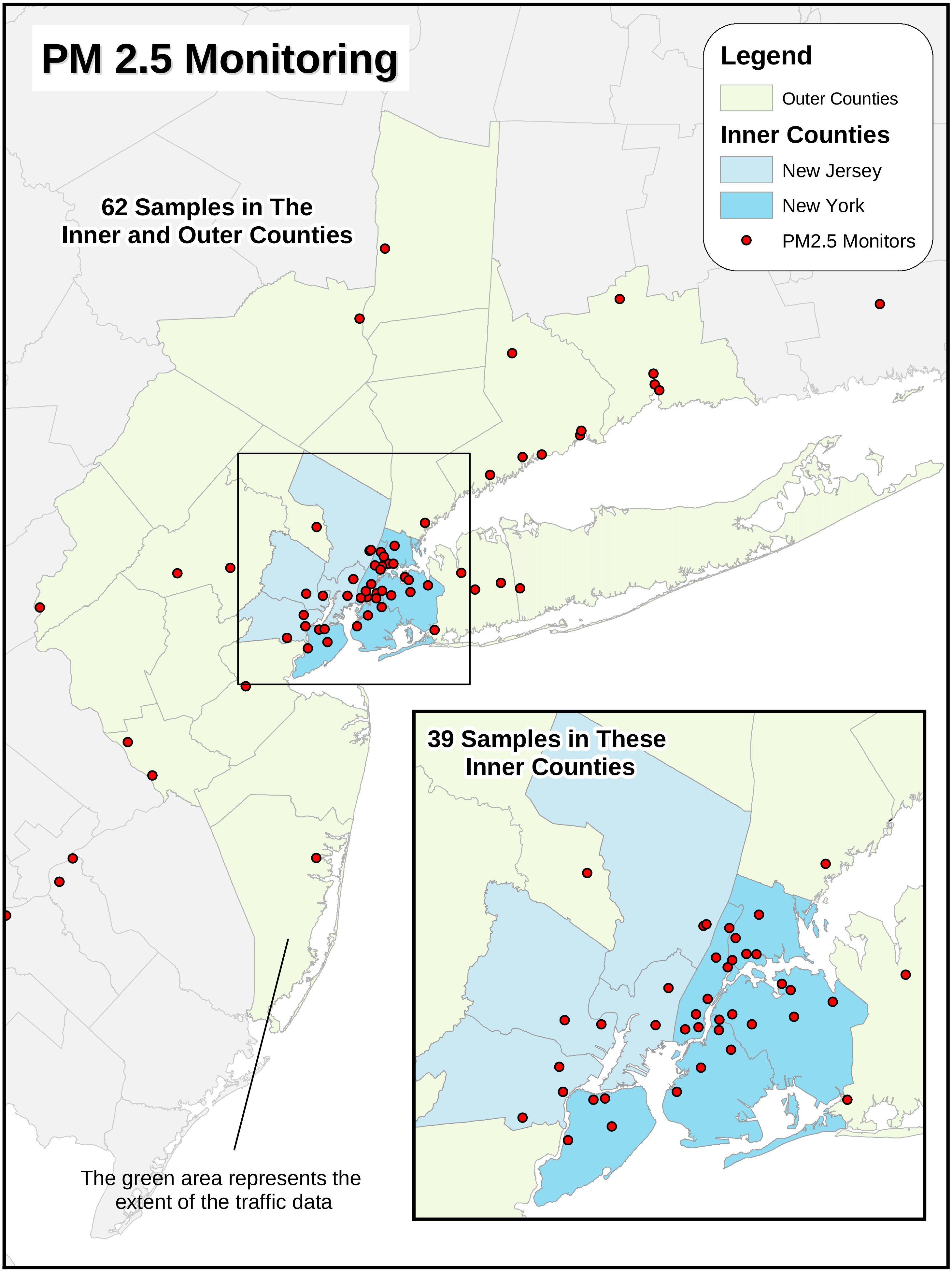
Example of buffering a monitor and clipping census block groups (1000m)



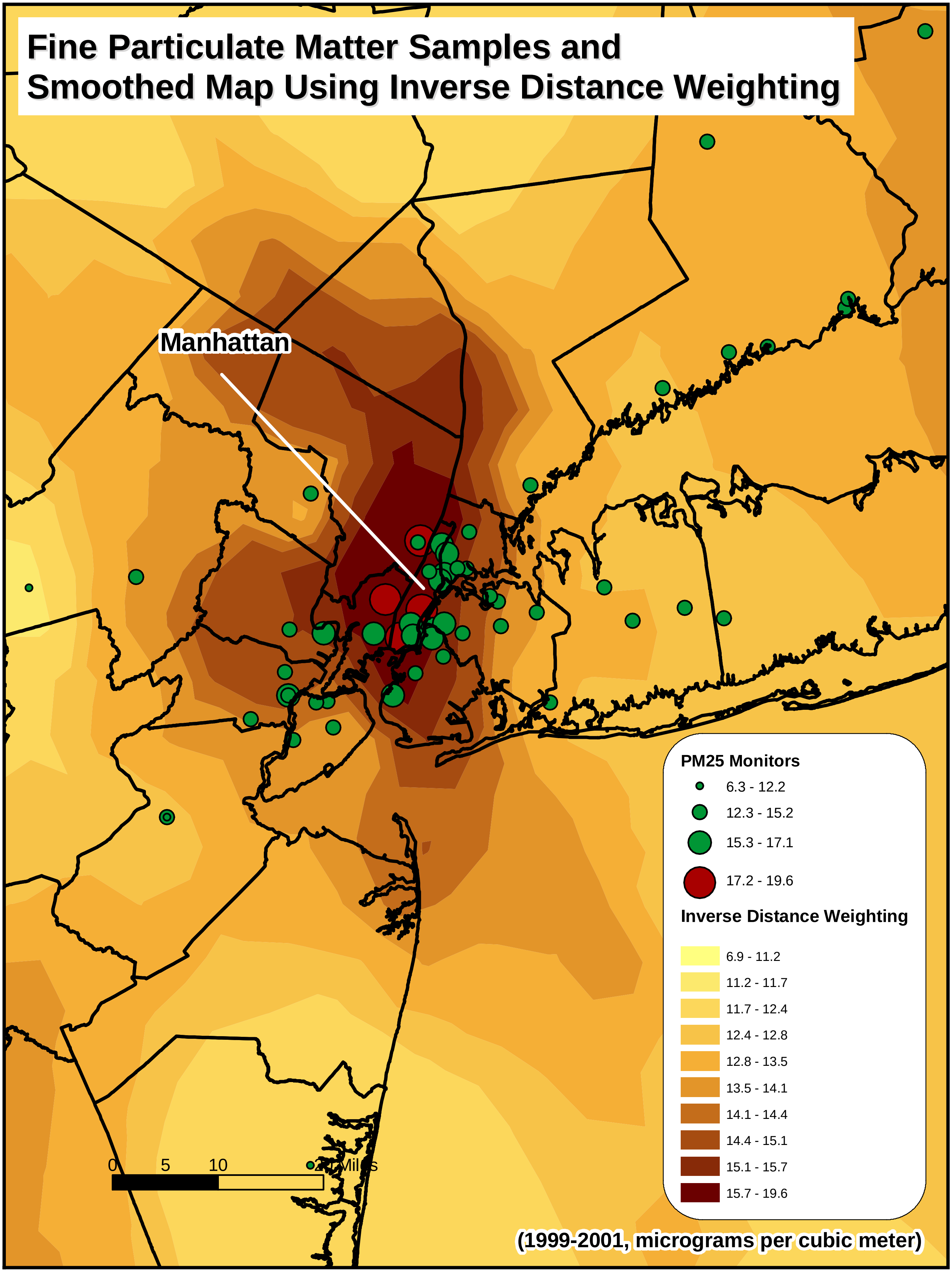
N

Prepared for the University of Southern California, Department of Preventive Medicine by ZevRoss Spatial Analysis

PM 2.5 Monitoring



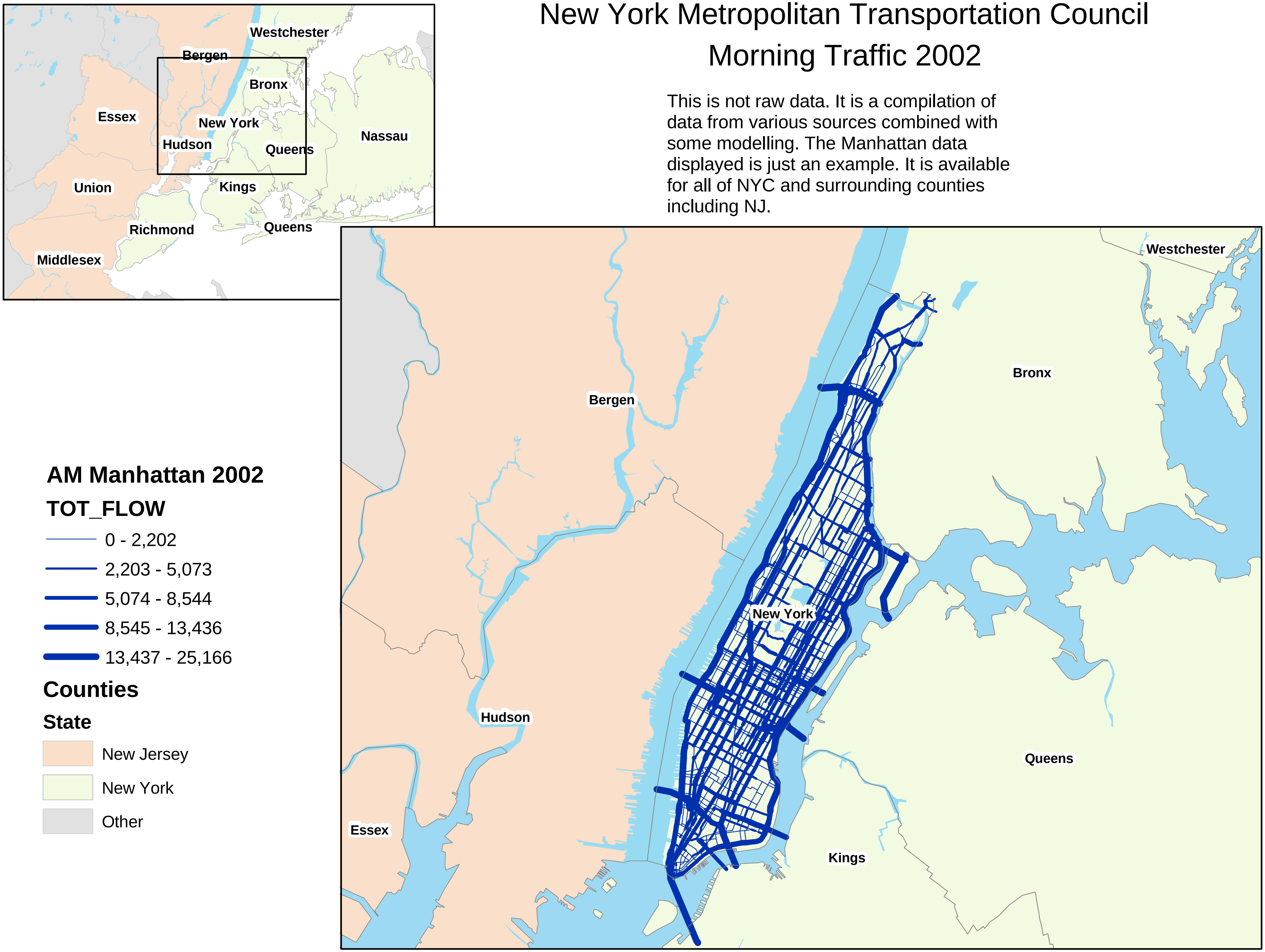
Fine Particulate Matter Samples and Smoothed Map Using Inverse Distance Weighting



(1999-2001, micrograms per cubic meter)

New York Metropolitan Transportation Council
Morning Traffic 2002

This is not raw data. It is a compilation of data from various sources combined with some modelling. The Manhattan data displayed is just an example. It is available for all of NYC and surrounding counties including NJ.



AM Manhattan 2002

TOT_FLOW

- 0 - 2,202
- 2,203 - 5,073
- 5,074 - 8,544
- 8,545 - 13,436
- 13,437 - 25,166

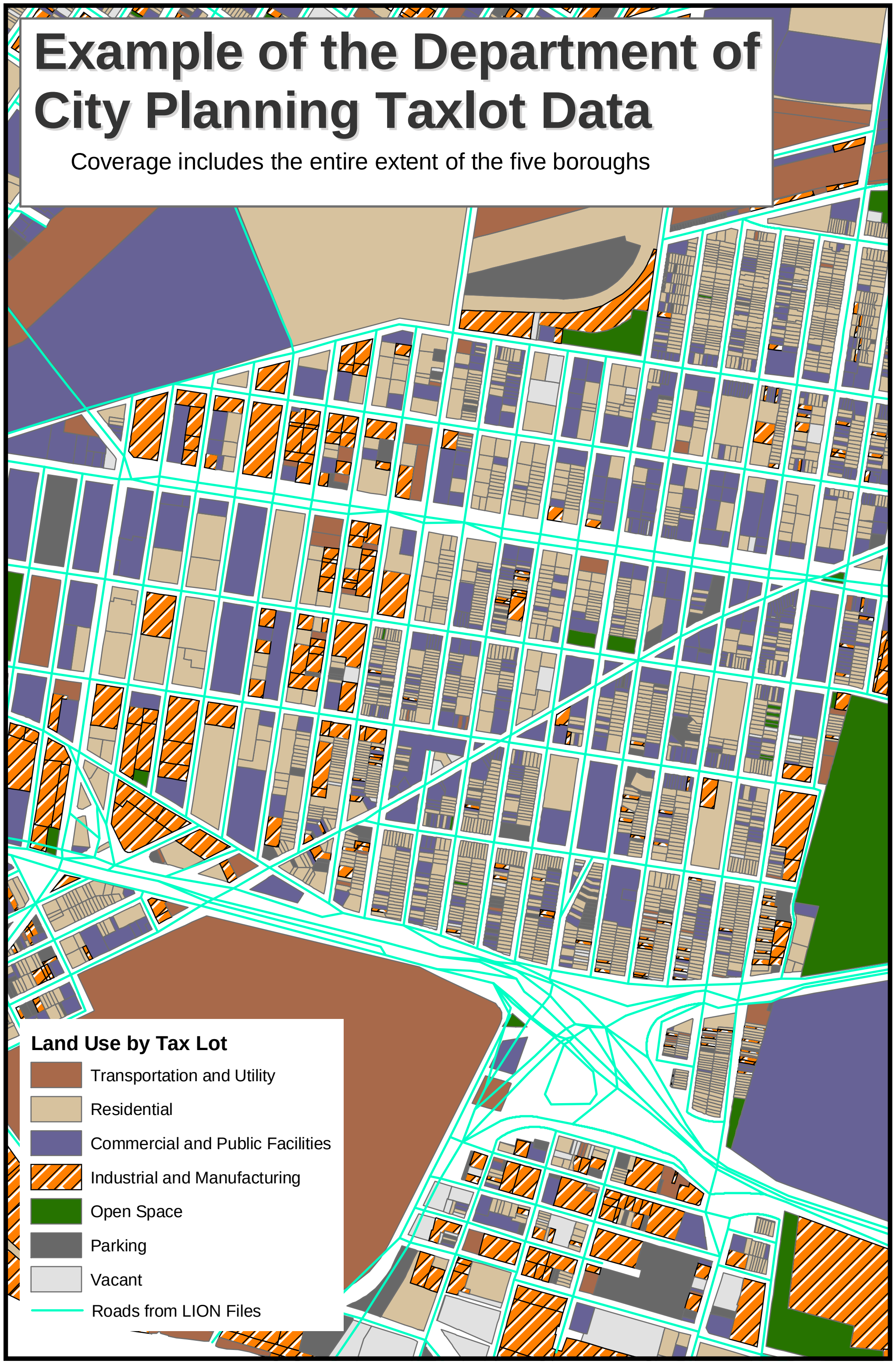
Counties

State

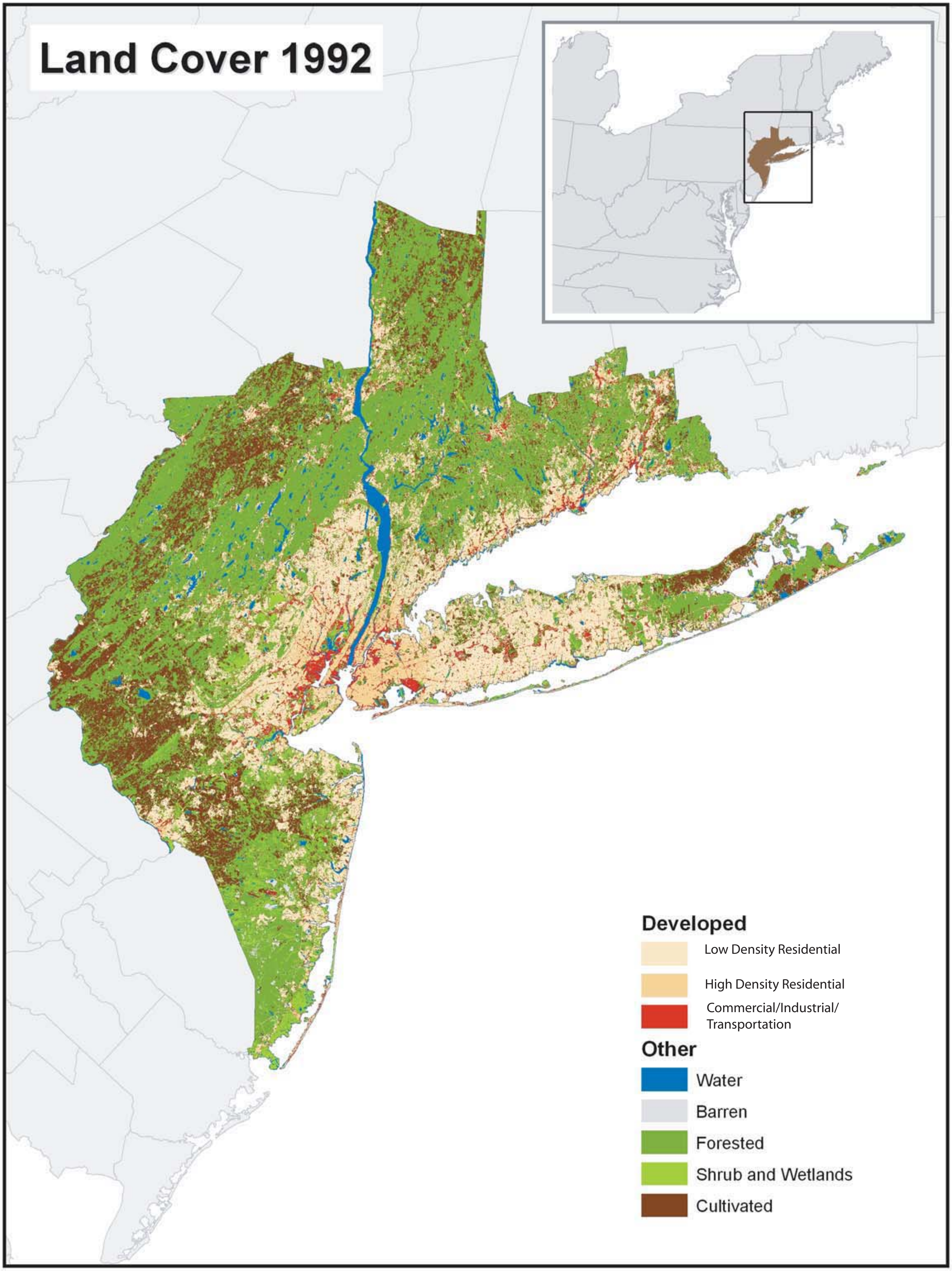
🟧	New Jersey
🟩	New York
⬜	Other

Example of the Department of City Planning Taxlot Data

Coverage includes the entire extent of the five boroughs



Land Use by Tax Lot
Transportation and Utility
Residential
Commercial and Public Facilities
Industrial and Manufacturing
Open Space
Parking
Vacant
Roads from LION Files


Land Cover 1992
Developed
Low Density Residential
High Density Residential
Commercial/Industrial/
Transportation
Other
Water
Barren
Forested
Shrub and Wetlands
Cultivated