Supplementary Information for

Exposure Density and Neighborhood Disparities in COVID-19 Infection Risk

Boyeong Hong, Bartosz J. Bonczak, Arpit Gupta, Lorna E. Thorpe, and Contantine E. Kontokosta

Corresponding Author: Constantine E. Kontokosta
E-mail: ckontokosta@nyu.edu

This PDF file includes:

- Figs. S1 to S5
- Tables S1 to S5
- Supplementary text
Fig. S1. Visualization of the data processing workflow starting from land use rasterization on a 1m x 1m grid (left), spatial join of mobility data with land use information by mapping activity locations on the raster (center) and aggregation of hourly activity by type on each 250m x 250m grid (right).
Fig. S2. Neighborhood social distancing metrics. Decrease of neighborhood activities (%) at non-residential buildings and outdoor land uses. Upper: Percentage change in neighborhood activity by zip code. Lower: Time series plots representing activity change over the time by zip code, with a Citywide moving average reference. Overall, we observe an approximately 20% decrease in neighborhood activities outside of residential buildings after the “stay-at-home” order.
Fig. S3. Neighborhood activity by land use type by 250m grid cell before and after the stay-at-home order: a) Citywide average neighborhood activity volume change b) Citywide average neighborhood activity proportion change.
Fig. S4. Hierarchical clustering tree diagrams by complete, average, and Ward’s linkage methods. We selected the Ward’s linkage method to minimize a within-group variance while maximizing between-group variance.
| Dataset               | Time range          | Resolution (spatial/temporal) | Source and description                                                                                                                                 |
|-----------------------|---------------------|--------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| Mobility data         | 2020-02-01 ~ 2020-04-30 | (X,Y)/second                | Ping geolocations provided by VenPath, Inc.                                                                                                            |
| NYC Primary Land Use Tax Lot Output (PLUTO) | updated 2020-02-24 | Parcel/NaN                  | Land use and building type information provided by the NYC Department of City Planning                                                                 |
| NYC Building Footprints | updated 2020-07-06 | Footprint/NaN                | Perimeter outlines of more than 1 million buildings in NYC provided by the Department of Information Technology & Telecommunications (DoITT). |
| Road Network Data (LION) | updated 2020-04-28 | Street segment/NaN           | Single line street base map provided by the NYC Department of City Planning                                                                       |
| NYC COVID-19 data     | 2020-04-01 ~ 2020-06-04 | Zip code/daily              | COVID-19 confirmed cases, deaths, and positivity rates provided by the NYC Department of Health and Mental Hygiene, retrieved from https://github.com/nychealth/coronavirus-data |
| American Community Survey (ACS) | 2018 5-year estimates | Zip code/NaN                | Neighborhood demographic and socioeconomic characteristics from the U.S. Census Bureau                                                              |
| NYC Hospital locations | updated 2017-09-08 | (X,Y)/NaN                    | Hospitals affiliated with the NYC Health and Hospital Corporation and public hospital system                                                           |
| Nursing home data     | updated 2020-05-24 | (X,Y)/NaN                    | Nursing home information, including the number of beds and occupancy, provided by the Centers for Disease Control's National Healthcare Safety Network system |
| Indicator      | Description                                              |
|----------------|-----------------------------------------------------------|
| Case rate      | Number of confirmed cases per 100,000 population by zip code |
| Death rate     | Number of confirmed deaths per 100,000 population by zip code |
| Positivity rate| Percentage of positive tests using a polymerase chain reaction (PCR) test |
| Deaths per case| Number of deaths per confirmed cases                      |
Table S3. Summary statistics of input variables

| Variable                        | Description                                           | Statistics   |
|---------------------------------|-------------------------------------------------------|--------------|
| Exposure density                | % change in activities outside of residential buildings | -0.18 (0.20) |
| Neighborhood cluster groups     | Dummy variables by cluster                             | -            |
| White                           | % non-Hispanic White population                        | 0.47 (0.27)  |
| Black                           | % Black population                                     | 0.21 (0.24)  |
| Asian                           | % Asian population                                     | 0.15 (0.14)  |
| Hispanic                        | % Hispanic population                                  | 0.26 (0.20)  |
| Age group 25-34                 | % of population 25-34 years old                       | 0.18 (0.06)  |
| Age group over 65               | % of population over 65 years old                     | 0.14 (0.05)  |
| Household size                  | Average household size in zip code                     | 2.64 (0.50)  |
| Household with kids             | % of households with kids under 18                     | 0.25 (0.06)  |
| Employees working from home     | % of employees working from home                       | 0.04 (0.03)  |
| Education attainment            | % of population with Bachelor’s degree                 | 0.23 (0.10)  |
| No health insurance             | % of households without health insurance               | 0.08 (0.04)  |
| Public health insurance         | % of households with public health insurance           | 0.39 (0.14)  |
| Commute time                    | Average commute time (minutes)                         | 40.76 (7.12) |
| Median income                   | Median household income in zip code                    | $74,000 ($37,000) |
| Unemployment rate               | % of labor force unemployed                            | 0.07 (0.03)  |
| Owner occupied units            | % of housing units occupied by owner                   | 0.37 (0.22)  |
| One or two family home          | % of one or two family home units                      | 0.30 (0.31)  |
| Public housing                  | % of public housing units                              | 0.05 (0.08)  |
| Residential area                | % of residential building area                         | 0.65 (0.20)  |
| Office area                     | % of office building area                              | 0.09 (0.15)  |
| Commercial area                 | % of commercial building area                          | 0.28 (0.17)  |
| Hospital                        | Whether a hospital(s) is located in zip code (yes=1)  | 0.21 (0.41)  |
| Nursing home                    | Number of occupied nursing home beds                  | 507 (837)    |

Standard deviations in parentheses.
Table S4. Multivariate regression model results based on the continuous exposure density variable

| Feature                                      | Model 1: case rate | Model 2: death rate | Model 3: positivity rate | Model 4: deaths per case |
|----------------------------------------------|--------------------|---------------------|--------------------------|-------------------------|
| Num of obs.                                | 177                | 177                 | 177                      | 177                     |
| F-stats.                                   | 45.25              | 23.28               | 71.95                    | 15.79                   |
| Prob > F = 0                               | 0.73               | 0.58                | 0.81                     | 0.49                    |
| Intercept                                  | 6.7820(0.170)***   | 2.6301(0.384)***    | 2.2074(0.113)***         | -3.7314(0.235)***       |
| Exposure density change                    | 0.0023(0.001)***   | 0.0005(0.003)       | 0.0023(0.001)**          | -0.0018(0.002)          |
| % Black                                    | 0.0056(0.001)***   | 0.0089(0.002)***    | 0.0046(0.001)***         | 0.0024(0.001)**         |
| % Hispanic                                 | 0.0107(0.001)***   | 0.0062(0.003)**     | 0.0058(0.001)***         | -0.0048(0.002)**        |
| % units occupied by owner                  | 0.0039(0.001)***   | -0.0038(0.003)      | 0.0033(0.001)***         | -0.0082(0.002)**        |
| % household with kids                      | 0.0109(0.003)***   | 0.0266(0.008)***    | 0.0117(0.002)***         | 0.0134(0.005)***        |
| % employees working from home              | -0.0294(0.008)**   | -0.0055(0.018)      | -0.0227(0.005)**         | 0.0156(0.011)           |
| Num of occupied nursing home beds per 100 people | 0.0405(0.011)*** | 0.0896(0.024)***   | 0.0125(0.007)*           | 0.0554(0.015)**        |
| % household without health insurance       | 0.0002(0.007)      | 0.0575(0.015)***   | 0.0153(0.005)***         | 0.0461(0.009)***        |
| Age group over 65                          | 0.0144(0.005)**    | 0.0705(0.011)***   | 0.0073(0.003)**          | 0.0440(0.007)***        |
| % Public housing area                      | -0.0045(0.003)     | 0.0068(0.006)      | -0.0015(0.002)           | 0.0094(0.004)**        |

Standard errors in parentheses. *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.10
Supplementary Information Text

Sensitivity Tests based on Rasterization Spatial Resolutions

Our land use classification data processing is based on a 1-meter × 1-meter grid. This granularity is intended to capture variations in land use in a high density urbanized area. However, we acknowledge that the horizontal accuracy of geolocation data has uncertainty caused by differences in device hardware and configuration, signal strength, or movement. The accuracy of each geolocated point is provided in the dataset as reported by the device and represents a 68% confidence radius. For the analyzed data, the mean reported horizontal accuracy is 52.6 meters and the median horizontal accuracy is 16.0 meters. Although we make the simplifying assumption that the point location provided in the data for each ping is accurate for the purposes of assigning a land use classification to a particular device, we also conducted a sensitivity test using different rasterization grid resolutions as a robustness check for our results.

Given the horizontal accuracy stated above, we re-run our analysis using two additional spatial resolutions - a 20-meter grid and 50-meter grid - to generate land use raster images. We create these grids and compute the land use type distribution by residential, non-residential, and outdoor areas, respectively, in each grid cell and assign the predominant land use type to the entire grid cell. Each geolocation data point from the mobility dataset is then assigned the activity type of that raster cell through a spatial join. We count the hourly number of unique devices in each neighborhood level 250-meter grid cell by the corresponding land use category associated with either the 20-meter grid raster or the 50-meter grid raster.

As described in the Methods section, we compute exposure density and estimate the behavioral response to social distancing by neighborhood, conduct hierarchical agglomerative clustering based on changes in neighborhood exposure density before and after the stay-at-home order (number of clusters, n = 5), and re-run the multivariate regression models based on the output cluster groups for each land use grid size. Figure S5a visualizes the spatial patterns of the average change in neighborhood exposure density before and after the stay-at-home order for the 1-meter grid, the 20-meter grid, and the 50-meter grid approaches. Overall, the spatial distribution and the magnitude of the exposure density change follow similar patterns, regardless of the underlying land use resolution used, with the mean difference of 0.018 and 0.025 for the 20-meter and 50-meter land use raster, respectively. We test the underlying distributions of the calculated exposure density for each land use resolution using the one-way analysis of variance (ANOVA) and find no statistically significant differences between the means resulting from the three approaches (F-statistic: 0.54; p-value: 0.59).

The differences in the agglomerative clustering results based on neighborhood activity change patterns are presented in figure S5b. By reducing the land use resolution to 20- or 50-meters, one can observe that the clustering algorithm does not capture as much of the variation in neighborhood activity patterns, which might suggest a greater reliance on underlying land use information than on mobility features. It is worth noting that regardless of the higher spatial dispersion, using a higher resolution land use raster results in greater between group variance of the clusters based on the calculated exposure density as shown in figure S5c. Despite the differences caused by applying rasters of varying resolutions, the overall clustering pattern holds with the cores of the major neighborhood groups being properly identified, such as Manhattan below 59th street forming group 1 (yellow); the Upper East Side, the Upper West Side, parts of Harlem, the East Village and Williamsburg areas forming group 2 (blue); and the predominantly residential areas of Staten Island, South Brooklyn, and Queens classified as group 5 (red).

The results of the multivariate regression positivity rate model are shown in table S5. Overall model goodness-of-fit, based on reported $R^2$ values, is consistent across land use raster resolutions. The neighborhood group level coefficients exhibit a reduction in magnitude as the land use raster grid size increases. The group variables maintain statistical significance across the model specifications, with the exception of two groups in the 50-meter grid model. The 1-meter resolution grid model is the preferred approach given the density and proximate variation in land uses in the study area, and the 20-meter grid, which is larger than the median horizontal location radius for the analyzed data, demonstrates the robustness of the exposure density methodology.
Fig. S5. Geolocation land use raster sensitivity results comparing 1-meter grid, 20-meter grid, and 50-meter grid resolutions. a) Neighborhood exposure density change by zip code. b) Agglomerative clustering results. c) Scatter plot of exposure density versus the log-transformed cumulative COVID-19 positivity rates through June 4th, 2020. Colors represent individual clusters.
Table S5. Geolocation land use raster sensitivity results: Multivariate regression positivity rate model results for model specifications using 1-meter grid, 20-meter grid, and 50-meter grid resolutions.

| Feature                                      | 1-meter grid        | 20-meter grid      | 50-meter grid      |
|----------------------------------------------|---------------------|--------------------|--------------------|
|                                              | Num of obs.=177     | Num of obs.=177    | Num of obs.=177    |
|                                              | F-stats.=53.26      | F-stats.= 52.01    | F-stats.= 40.09    |
|                                              | ^Prob>F=0=0.83       | ^Prob>F=0=0.83     | ^Prob>F=0=0.82     |
|                                              | R-squared=0.83      | R-squared=0.83     | R-squared=0.82     |
| Intercept                                   | 2.359(0.116)***     | 2.302 (0.109)***   | 2.280 (0.128)***   |
| Group “outflow-mixed” and “outflow-residential” | -0.443(0.091)***   | -0.342 (0.081)***  | -0.308 (0.106)***  |
| Group “stable-outflow”                       | -0.228(0.096)**     | -0.300 (0.112)***  | -0.094 (0.114)     |
| Group “shelter-in-place”                     | -0.130(0.078)*      | -0.189 (0.075)**   | -0.052 (0.097)     |
| % Black                                      | 0.004(0.001)***     | 0.004(0.001)***    | 0.004(0.001)***    |
| % Hispanic                                   | 0.005(0.001)***     | 0.005(0.001)***    | 0.005(0.001)***    |
| % units occupied by owner                    | 0.003(0.001)***     | 0.013 (0.002)**    | 0.003 (0.001)**    |
| % household with kids                        | 0.013(0.002)***     | 0.013(0.002)**     | 0.013(0.002)**     |
| % employees working from home                | -0.016(0.005)***    | -0.018 (0.005)**   | -0.018 (0.005)**   |
| Num of occupied nursing home beds per 100 people | 0.008 (0.007)     | 1.312 (0.695)*     | 0.933 (0.711)      |
| % household without health insurance          | -0.003(0.007)       | 0.002 (0.006)      | 0.007 (0.008)      |
| Insurance × group effect 1                   | 0.046(0.012)***     | 0.045 (0.009)***   | 0.030 (0.011)**    |
| Insurance × group effect 2                   | 0.021(0.010)**      | 0.024 (0.012)**    | 0.007 (0.011)      |
| Insurance × group effect 3                   | 0.018(0.009)**      | 0.022 (0.008)**    | 0.008 (0.010)      |
| Age group over 65                            | 0.008(0.003)***     | 0.008(0.003)**     | 0.008(0.003)**     |
| % Public housing area                        | -0.002(0.002)       | -0.002(0.002)      | -0.002(0.002)      |
| AIC                                          | -151.9              | -148.40            | -140.0             |
| BIC                                          | -101.1              | -97.58             | -89.14             |

Standard errors in parentheses. *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.10