Multivariate analysis of the influence between building design and energy performance, socio-demographic metrics, and the intra-urban environment

M Llaguno-Munitxà¹, X Shu², B Mistry²

¹UCLouvain SST | LOCI Place du Levant 1, 1348 Louvain-la-Neuve, Belgium.
²Northeastern University | 360 Huntington Ave, Boston, MA 02115, United States.

maider.llaguno@uclouvain.be

Abstract. Through advancements of direct and remote sensing technologies, we have recently learned that urban microclimate and air quality gradients can often be more significant than city to rural differences. However, the urban design parameters that are most critical to improve environmental health and reduce building energy consumption, are yet to be identified. This research makes use of air quality datasets collected through a dense stationary sensing network in New York City, remote sensing datasets for land surface temperature and normalized difference vegetation index, building usage metrics, building and urban design metrics, and socio-demographic datasets including population and health metrics. Through a neighbourhood-scale footprint-based regression analysis, the correlation between the air quality, land surface temperature, building usage and urban metrics has been studied. Highest correlations have been observed between air quality and land surface temperature and urban design and socio-economic metrics. The results show that building usage metrics such as the energy use intensity or electricity purchase, are mainly affected by building design characteristics. On the other hand, significant correlations have been observed between the urban design, socio-demographic and contaminant concentration gradients, addressing the critical role the planning and design of our cities plays in the environmental well-being of citizens.

1. Introduction

Driven by the insalubrious environmental conditions that threaten many global cities, initiatives to improve urban environmental health are becoming one of the main priorities of local governments and municipal authorities. Aiming to capture the intra-urban environmental gradients, in the last few years dense stationary and mobile sensing technologies have been deployed in several global metropolises [1] [2] [3]. Such datasets are enabling the identification of the neighbourhoods that are most affected by heat stress, poor air quality conditions, or environmental inequality as a whole [4] [5] [6]. Recent access to fine grain remote sensing datasets on land cover and surface temperature, in combination with detailed urban experimental environmental datasets, are also enabling the identification of the dependencies between urban environmental gradients, urban design and socio-demographic parameters, and building use performance metrics opening new possibilities to more accurately assess the environmental design and energetic performance of buildings.

Prior literature has reported that the temporal fluctuations in energy consumption demands are linked to regional weather patterns [7] and that spatial variations are often attributable to household typologies or building density metrics [8]. On the other hand, energy consumption datasets have also been identified as key indicators of standards of living or household income [9] [10]. At the city scale,
increased land surface temperatures and urban heat island intensity have also been associated to peaks in household energy consumption [11] - [13]. Recent literature has also reported that higher outdoor air contaminant concentrations are associated to increased household energy consumptions [14] [15]. However, the study of the intra-urban spatiotemporal energy consumption gradients remains yet to be further explored [16] [17]. Prior studies have mainly focused on inter-city or single building scale studies through the use of regional weather datasets, or city scale environmental metrics, failing to address the role intra-urban microclimate and air quality gradients play in building usage metrics. This research proposes a multivariate analysis where the influence between building design, usage, socio-demographic metrics, and intra-urban air quality gradients and land surface temperature will be studied. This paper makes use of the air quality datasets collected by the New York City’s Department of Health and Mental Hygiene’s: New York City Community Air Survey (NYCCAS). Air quality data was collected for 10 years (2008-2018) through ~150 stations distributed across the city. Remote sensing datasets from Landsat 8 on the Land Surface Temperature (LST) and Normalized Difference Vegetation Index (NDVI) have also been derived at 30x30m spatial resolutions and for 3 time periods of 16 days in 2017. NYC also makes extensive urban datasets available on the physical attributes of the city (building and street design characteristics), socio-demographic metrics (population and health indicators), as well as building use metrics (energy consumption or greenhouse gas emissions). These datasets will be used to explore the interrelations between air quality, LST, NDVI, urban design, socio-demographic and building usage metrics. Prior publications that have made use of the NYC energy consumption databases reported linkages between socio-demographic and energy consumption metrics [18], as well as between building typologies and energy consumption records [15]. However, in these studies detailed intra-urban environmental datasets were not taken into account. The main questions these research aims to address are the following: What role do local urban air quality and land surface temperature gradients play in the building usage metrics such as energy consumption? What are the linkages between building and city characteristics and energy consumption metrics? How can we make use of detailed intra-urban urban environmental data to better design our buildings and cities?

2. Methodology
This paper presents the results from a comparative multivariate analysis that utilizes the air quality, LST, NDVI, urban design, socio-demographic and building usage datasets. Overall, 55 parameters were collected from 6 main data sources; NYCCAS, Landsat, PLUTO, Census Open New York data, LL84: Energy and Water Data Disclosure, and American Community Survey (ACS). Following the data collection process, a data pre-processing was developed. The data formatting process was a critical step given the differences in spatial granularity and temporal resolution availability for the different datasets. Once the data was formatted, the data validation and parameter identification process was developed. From the original 55 parameters, 28 were featured based on a GIS location-based study. The features are classified according to four main categories as displayed in Table 1: i) Urban environment, ii) urban design, iii) socio-demographic parameters and iv) building usage metrics. The featured parameters for the urban design category include total building area Building$_{a}$, building footprint area Building$_{fa}$, total residential area Residential$_{a}$, total office area Office$_{a}$, mean building height Height$_{ma}$, total road length Road$_{a}$, total paving area Paving$_{a}$, mean ground elevation Elevation$_{a}$, total green area Grass$_{a}$ and street level tree count Tree$_{a}$. The socio-demographic category includes data on the total population Population$_{a}$, the total population over 65 years old Population$_{as5}$, total elementary school population Population$_{as}$, total high school population Population$_{ah}$, median income Income$_{med}$, total unemployment Unemployment$_{a}$, total asthma case hospitalizations Asthma$_{a}$, and total chronic respiratory disease hospitalizations Respiratory$_{dch}$. Finally, the building usage metrics include total energy use EnergyUse$_{e}$, energy use intensity $EUI$, total green-house gas emissions GHG$_{e}$, green-house gas emissions per square foot GHG, total electricity purchase Electricity$_{e}$, electricity purchase per square foot Electricity, total natural gas consumption NaturalGas$_{e}$, and natural gas consumption per square foot NaturalGas. Aside from the urban parameters, air quality data has been collected from the New York City Air Survey (NYCCAS) from 2008 to 2018. The NYCCAS air quality dataset is one of the most complete air quality datasets available today that includes data for Nitrogen Oxide (NO), Nitrogen Dioxide (NO$_2$), Particulate
Matter <2.5μm (PM₂.₅), Ozone (O₃), Elemental Carbon (EC), and Sulfur Dioxide (SO₂). Prior literature has reported high correlations > 0.9 between traffic related contaminants [19] [20], and thus solely the contaminant PM₂.₅ has been used in this study. The dataset for the LST and NDVI have been derived from Landsat 8 and for three time periods from June, August and February 2017 at 30 × 30m resolution.

### Table 1. Data sources.

| Category             | Sub-Category | Data source                                                                                                                                 |
|----------------------|--------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| Urban Environment    | Air Quality(1) | 1. The New York City Air Survey (NYCCAS) from 2008 to 2018 Parameters: \( \text{PM}_{2.5} \, \text{μg/m}^3 \). |
|                      |              | 2. Remote Sensing: Landsat 8. February 2017, June 2017, August 2017 Parameters: LST (°C). |
|                      | NDVI (1)     | 1. Normalized Difference Vegetation Index Parameters: NDVI (index)                                                                        |
| Urban Design         | Land Use (7)  | 1. New York City Open Data for Land Use 3ft resolution. Parameters: Tree, Grass, Earth, Water, Buildings, Road, Paving. |
|                      | Buildings (5) | 1. Building Footprints – NYC Open Data Parameters: Building, Footprints – NYC Department of Planning |
|                      | Streets (2)  | 1. NYC Street Centerline (CSCL) – NYC Open Data Parameters: Elevation, Street.                                                              |
| Socio-Demographic    | Socio-Economic (6) | 1. American Community Survey (ACS) by Neighborhood Tabulation Area (NTA), 2014-2018 – NYC Department of City Planning. Parameters: Population, Population₂, Population₃, Population₄, Population₅, Population₆, Population₇, Population₈, Population₉, Income, Unemployment. |
|                      | Health (2)   | 1. NYC Environment & Health Data Portal - Average annual 2012-2014 asthma-related hospitalizations among resident adults by NTA. 2. New York City Department of Health and Mental Hygiene. Number of Chronic Obstructive Pulmonary Disease (COPD) hospitalizations Parameters: Asthma₂, Asthma₃, Asthma₄, Asthma₅, Asthma₆, Asthma₇, Asthma₈, Asthma₉, Asthma₁₀. |
| Building Use Metrics | Consumption (3) | 1. 2018 Energy and Water Data Disclosure (Data Years 2012, 2014 & 2017) NYC Mayor’s Office of Sustainability Parameters: EUI (kBtu/ft²), EnergyUse (kBtu), Electricity (kBtu/ft²), NaturalGas. |
|                      | Emissions (1) | 1. 2018 Energy and Water Data Disclosure (Data Years 2012, 2014 & 2017) NYC Mayor’s Office of Sustainability Parameters: GHG (Metric Tons CO₂e/ft²), GHGe (Metric Tons CO₂e). |

The NYCCAS includes biweekly recordings of the contaminants. As reported in [2], the data from the different stations was collected during different weeks, and some stations were not active during periods of several months given maintenance requirements. In order to account for the missing data, the dataset for all collected years 2008-2018 was separated in meteorological seasons. The stations for which at least a 2-week data collection per season was performed were included in the calculations. For the Land Surface Temperature (LST) and Normalized Difference Vegetation Index (NDVI), datasets from three timeframes of 16 days each were collected for the months of January, June and August. The inter-correlation between the collected LST were computed for the studied timeframes and correlations of 0.66 Feb-June, 0.60 Feb-Aug, and 0.86 June-Aug were found. For the case of NDVI correlations of 0.71 for Feb-June, 0.79 for Feb-Aug, and 0.92 for June-Aug were observed. For PM₂.₅, correlations > 0.75 were observed between the studied years. Given the observed correlations, despite the timeframe differences between the collected datasets, it has been considered acceptable to develop a comparative study. Figure 1 displays the distribution of the NYCCAS air quality stations coloured based on the mean PM₂.₅ concentration values, and the gradient plot of the LST temperature deviation across the city. The paper will present the results using PM₂.₅ data for the years 2012, 2014 and 2017 given the availability for building usage data for the same years. Table 2 shows the statistical correlations between the Energy Usage Intensity (EUI) and Greenhouse Gas (GHG) emission metrics for all years. Over the studied years
the Borough Block Lot (BBL) property identifiers vary, thus the differences on the building usage correlations are also likely related to the differences in the building stock. The figures displayed in this paper include the results obtained utilizing the building usage data from 2017 given the increased data availability in comparison to the previous years for which 346861 entries were collected, and 19194 were kept after the data processing steps. The top and bottom 1% of buildings by EUI were omitted as outliers, and removal of duplicate entries was deployed.

Table 2. Building usage metric statistics for 2012, 2014 and 2017.

|        | EUI 2012 | EUI 2014 | EUI 2017 | GHG 2012 | GHG 2014 | GHG 2017 |
|--------|----------|----------|----------|-----------|-----------|-----------|
| EUI 2012 | 1        | 0.611    | 0.452    | 0.569     | 0.273     | 0.212     |
| EUI 2014 | 0.611    | 1        | 0.534    | 0.258     | 0.565     | 0.264     |
| EUI 2017 | 0.452    | 0.534    | 1        | 0.1722    | 0.230     | 0.450     |
| GHG 2012 | 0.569    | 0.258    | 0.1722   | 1         | 0.759     | 0.639     |
| GHG 2014 | 0.273    | 0.565    | 0.230    | 0.759     | 1         | 0.697     |
| GHG 2017 | 0.212    | 0.264    | 0.450    | 0.639     | 0.697     | 1         |

Figure 1. a) NYCCAS stations coloured based on the mean PM$_{2.5}$ concentration, b) Land Surface Temperature map of NYC for June in 2017.

Following the completion of the data compilation and processing steps, a comparative multivariate analysis was developed though a neighbourhood-scale footprint-based regression methodology as presented in [21]. The socio-demographic datasets are recorded at the Neighbourhood Tabular Area (NTA) which exceeds the building footprint scales. In these instances, a tabular intersection method (ArcGIS pro 2.7) has been deployed to map the local NTA level socio-demographic datasets to the PLUTO property scale following the Borough Tax Block & Lot (BBL) identifier. The building usage dataset on the other hand is reported at the building category PLUTO tax lot data as it is the case for the building design metrics. Once all datasets were mapped at the building property level, urban parameter metrics have been computed for each NYCCAS air quality station location, and for distinct urban footprint sizes. Urban footprint scales of $r = 100$ m, $r = 250$ m, $r = 500$ m and $r = 1000$ m have been computed (See Figure 2.a). The NYC energy dataset reports the building EUI and EnergyUse, GHG, Electricity, and NaturalGas information for about ~3% of the total number of buildings recorded by NYC Open Data for the case of 2017 for example. Figure 2.b display the location of buildings with available building use data for one of the NYCCAS stations and for a $r = 250$ m footprint. The building usage metrics were computed using the available mean building consumption and emission data contained within the footprint and assigning it to the Building$_{in}$. The second methodology presented in the paper includes a building-based regression analysis using the buildings we have building usage data for as anchor points. For this study, the LST was computed identifying the LST values contained within
the building footprint geometry (see Figure 2.c). For PM$_{2.5}$ on the other hand, given that we don’t have air quality data for all streets in the city, an outdoor air quality estimation model using a weighted averaged method was used and an estimated concentration value was assigned to each of the buildings. The available urban design metrics at the building scale where intersected and assigned to the building usage records.

![Figure 2. a) Building footprints r = 100m, 250m, 500m & 1000m for a NYCCAS station b) Footprint r = 250m with building use datasets marked in red for a NYCCAS station located in Bronx c) Building-based study showing the LST intersection for buildings with available energy data.](image)

3. Results
In order to identify the independent variables and understand the interdependences between the collected parameters for all categories, the Pearson correlation [22] between all urban parameters under study and for r = 100 m, r = 250 m, r = 500 m and r = 1000 m footprint sizes have been computed and the results obtained for r = 250 m are presented as illustration in Figure 3. Highest correlations are observed between urban design related parameters, and socio-demographic parameters. Through this study the independent variables and those with most significant correlations were identified which have been presented in Figure 3.b and 3.c. The two-health metrics Asthma$_{ch}$ and Respiratory$_{cdh}$ were highly correlated, however they were kept in the following studies given that they were the only two health metrics we had access to.

![Figure 3. a) Pearson correlation matrix for r = 250 m b) Pearson correlation bar plot for PM$_{2.5}$ as the dependent variable c) Pearson correlation bar plot for LST as the dependent variable.](image)
As shown in Figure 3b highest correlations are observed between PM$_{2.5}$ and the urban design parameters such as Height$_{bm}$, the Building$_{ta}$, or Office$_{ta}$. Significant negative correlations are also observed between greenery related parameters that is Tree$_{ta}$, Grass$_{ta}$, and NDVI. Within the socio-demographic category, the highest correlations with PM$_{2.5}$ have been observed with metrics associated to Population, Income$_{med}$ and Unemployment. These findings reveal the underlying socio-environmental inequality in the city. The health metric parameters on the other hand show low correlations with the PM$_{2.5}$ dataset for all studied footprints. This is most likely related to the limitations of the health data as well as the number of NYCCAS stations that were kept after the data cleaning process. In order to further explore this point, we identified the 10% highest PM$_{2.5}$ air contaminant concentration stations in the city and compared them against the health metrics to see that the neighbourhoods with highest hospitalizations are amongst the ones exposed to the worst air quality. For the parameters grouped within the building usage category, the correlations with PM$_{2.5}$ concentrations appear to be mostly attributable to building properties such as the Building$_{ta}$. Figure 3.c displays the correlations for LST as the dependent variable. As it was expected, LST shows negative correlations against greenery related parameters Tree$_{ta}$, Grass$_{ta}$ and NDVI. Strong positive correlations are observed against urban parameters that relate with surface imperviousness such as Roads$_{ta}$, Paving$_{ta}$, or Buildings$_{ta}$. The LST and energy usage intensity show negative correlations which can be attributed to the predominance in heating loads that characterize the energy consumption statistics of New York City. Given the limited number of NYCCAS stations and availability of building usage data contained within the footprints, an alternative study has been developed through a building-based regression analysis where the building specific usage metrics: EnergyUse$_{t}$, GHGe$_{t}$, Electricity$_{p}$ and NaturalGas$_{c}$, and design metrics: Building$_{ta}$, Building$_{fa}$ and Height$_{bm}$ have been included. The PM$_{2.5}$ concentration has been obtained deploying a distance weighted PM$_{2.5}$ estimation method for each building containing energy use data. Figure 4a displays the buildings with available building usage data coloured based on their assigned PM$_{2.5}$ concentration. For these studies, the building usage data has also been studied separating the different building types such as commercial, residential or industrial buildings amongst others.

For the building-based study, the available data has been studied as a group, and separated in functional types: residential, commercial, mixed-use and industrial. In Figure 4.b the studied metrics for solely commercial buildings are displayed. The observed correlations vary between the studied typologies. Highest correlations between building design related parameters and building usage parameters have been observed for commercial building typologies, followed by residential building typologies. For all cases, significant correlations have been observed amongst the building use parameters and building geometry characteristics such as Building$_{ta}$ or Height$_{bm}$. Given that the datasets

---

**Figure 4.** a) Building usage data buildings coloured by PM$_{2.5}$ estimation b) Building-based regression for commercial buildings with available building use metrics.
for the building usage parameters presented in the plot include the total consumption metrics, the correlation with building design characteristics was expected. On the other hand, low correlations have been observed between building use metrics and PM$_{2.5}$ concentrations. The differences between the neighbourhood scale footprint-based study and the building-based study can be partially attributable to the PM$_{2.5}$ estimation model used in the building-based study. Most buildings have their nearest PM$_{2.5}$ stations over 500 m away, and so the weighted PM$_{2.5}$ concentration estimation may not represent accurately the gradients that characterise the urban environment. Overall negative correlations are observed between LST and Height, and Building, parameters. Subtle positive correlations have been observed between LST and building compactness related parameters such as Lotfront or Lotdepth. On the other hand, distinct correlation trends have been observed between LST and building usage metrics at the building-based study, as opposed to the results observed for the neighbourhood scale footprint-based study. The differences observed between the results obtained through the developed methodologies reveal an underlaying influence of the urban environment on the building usage metrics, suggesting that the neighbourhood scale microclimate plays a role in the building consumption metrics.

4. Conclusions

Urban datasets of the city of New York for urban design and socio-demographic parameters, building usage metrics, and local air contaminant concentrations and land surface temperature have been compiled. Comparative multivariate studies have been developed to understand the interrelations between the studied parameters. The results obtained for the years 2012, 2014 and 2017 have been presented in the paper. Through a footprint-based analysis using 4 distinct footprint scales: r = 100 m, r = 250 m, r = 500 m & r = 1000m, the urban metrics have been computed for each air quality station location and the correlations between the distinct urban features have been studied to evaluate the role intra-urban air quality conditions, land surface temperature, and urban design and socio-demographic parameters play in the building usage metrics. A building-based methodology focusing on the buildings with available building usage data has also been presented. The results show that highest correlations between building usage metrics such as the energy use intensity, electricity purchase or greenhouse emissions, are mainly driven by building design characteristics. Overall, the correlations between building usage metrics and PM$_{2.5}$ appear to be mostly attributable to the total building area and volume metrics. The negative correlations observed between building usage metrics and LST can be attributed to the predominance in heating loads that characterize the energy consumption statistics of New York City. Significant correlations have also been observed between building usage and socio-demographic metrics such as population or economic status. Low correlations have been observed between the health-related and building usage metrics. These findings contribute to the understanding of the critical role urban planning and design parameters play in the environmental well-being of citizens, as well as in the building consumption patterns in cities.

References

[1] Environmental Defense Fund 2020 “The breathe London blueprint: how cities can use hyperlocal air pollution monitoring to support their clean air goals.” https://www.globalcleanair.org/files/2021/02/EDF-Europe-BreatheLondon_Blueprint-guide.pdf
[2] The New York City Community Air Survey: neighbourhood air quality 2018 www1.nyc.gov/assets/doh/downloads/pdf/environmental/comm-air-survey-08-16.pdf
[3] Chen L J, Hsu W, Cheng M and Lee H C 2016 “A location-aware sensing system for participatory PM2.5 monitoring MobiSys” Companion Proc. 14th Annual Int. Conf. on Mobile Systems, Applications, and Services Companion 98–98.
[4] Environmental Defense Fund 2019 “Making the invisible visible: a guide for mapping hyperlocal air pollution to drive clean air action”. https://www.edf.org/sites/default/files/content/making-the-invisible-visible.pdf
[5] Harlam S L, Chowell G, Yang S, Morales E, Petitti D B, Ruddell B L and Ruddell D M 2014 “Heat-related deaths in hot cities: estimating human tolerance to high temperatures” International Journal of Environmental Research Letters and Public Health 11(3): 3304-3326.
[6] Wouters H, De Ridder K, Poelmans L, Willems P, Brouwers J, Hosseinzadehtalaei P, Tabari H, Vanden Broucke S, Van Lipzig N P M and Demuzere H 2017 “Heat stress increase under climate change twice as large in cities as in rural areas: a study for a densely populated midlatitude maritime region” Geophysical Research Letters, 44:17:8997-9007.

[7] Ortiz L E, Gutierrez E, Gonzalez J and Arend M 2016 “Forecasting building energy demands With a coupled weather-building energy model in a dense urban environment” Journal of Solar Energy Engineering 139.

[8] Ewing R and Rong F 2008 “The impact of urban form on U.S. residential energy use” Housing Policy Debate, 19:1, 1-30.

[9] Joyeux R and Ripple R D 2007 “Household energy consumption versus income and relative standard of living: a panel approach” Energy Policy, 35(1), 50–60.

[10] Mashhoodi B, Stead D and Van Timmeren A 2019 “Spatial homogeneity and heterogeneity of energy poverty: a neglected dimension” Annals of GIS, 25(1), 19–31.

[11] Santamouris M, Cartalis C, Synnefa A and Kolokotsa D 2015 “On the impact of urban heat island and global warming on the power demand and electricity consumption of buildings—a review” Energy and Buildings, 98, 119–124.

[12] Santamouris M, Papanikolaou N, Livada I, Koronakis I, Georgakis C and Argiriou A 2011 “On the impact of urban climate on the energy consumption of buildings” Solar Energy 70-3.

[13] Kumari P, Kapur S, Garg V and Kumar K 2020. “Effect of surface temperature on energy consumption in a calibrated building: a case study of Delhi” Climate 8(6) 71.

[14] You S, Eoh K G and Tong Y W 2017 “Variation of household electricity consumption and potential impact of outdoor PM2.5 concentration” Applied Energy 188:475-484.

[15] He P, Liang J, Qiu Y, Li Q and Xing B 2020 “Increase in domestic electricity consumption from particulate air pollution” Nature Energy 5 985-995.

[16] Mashhoodi B, Stead D and Van Timmeren A 2020 “Land surface temperature and households’ energy consumption: who is affected and where?” Applied Geography 114.

[17] Ma J and Cheng C P J 2016 “Estimation of the building energy use intensity in the urban scale by integrating GIS and big data technology” Applied Energy 183 182-192.

[18] Kontokosta C E 2012. “Local law 84 energy benchmarking data: report to the New York City mayor’s office of long term planning and sustainability”. http://www.nyc.gov/html/gbee/downloads/pdf/ll84_kontoska_report.pdf

[19] Matte T D, Ross Z, Kheirbek I, Eisl H, Johnson S, Gorczynski J E, Kass D, Markowitz S, Pezeshki G and Clougherty J E 2013 “Monitoring intraurban spatial patterns of multiple combustion air pollutants in New York City: design and implementation” J Expo Sci Epid 23 223–31.

[20] Llaguno-Munitxa M and Bou-Zeid E (2021) “Sensing Environmental Neighborhoods. In: Yuan P.F., Yao J., Yan C., Wang X., Leach N. (eds) Proceedings of the 2020 DigitalFUTURES. CDRF 2020, Springer, Singapore. https://doi.org/10.1007/978-981-33-4400-6_12

[21] Llaguno-Munitxa M and Bou-Zeid E (2020) “Environmental neighborhoods and their spatial extent” Environmental Research Letters. 15, 074034.

[22] Pearson K 1895 “Notes on regression and inheritance in the case of two parents” Proc. Royal . Soc. Lond. 58 240