# A DATA-DRIVEN AND PHYSICS-BASED FRAMEWORK FOR INTEGRATED ENERGY-AIR QUALITY (iE-AQ) MODELING


## Abstract

To date, limited work has been implemented to integrate energy and air-quality models in a unified system of built environment design. This limitation can be more critical for performance evaluation of naturally ventilated buildings in which understanding the trade-offs and synergies between energy-saving goals and indoor air quality is of primary importance. The main objective of this research is to develop a framework for integrated Energy-Air Quality (iE-AQ) modeling to support data-informed decisions for reducing human health risks and energy consumption for the built environment. This framework hybridizes data-driven and physics-based platforms, and brings the power of artificial intelligence (AI) techniques into the conventional simulation workflows to enable a more reliable and efficient approach. The proposed framework identified spatiotemporal factors that explain outdoor air quality variation across urban areas and localized outdoor air pollution, herein $PM_{2.5}$, through the Land Use Regression (LUR) based using Gradient Boosting Machine (GBM) approach (LUGB). The proposed framework was tested on the prototype large-size office buildings provided by the U.S. Department of Energy (DOE) across the City of Chicago. The obtained $R^2$=0.71 from LUGB suggests the power of this approach compared with the traditional LUR model with multiple linear regression (MLR) ($R^2$=0.43) for localizing hourly outdoor $PM_{2.5}$ concentrations. The variations of energy-saving potentials were obtained 6.4% to 15.6%, showing the significance of the proposed approach for evaluation of naturally ventilated buildings by generalizing outdoor conditions, not compromising human health. This research has the potential to aid designers, engineers, planners, and policymakers with a better awareness of the existing profiles of urban air quality variations across urban districts to achieve sustainable built environment goals.



**Author**

Mehdi Ashayeri
*Illinois Institute of Technology*



**Keywords**

Natural ventilation, energy saving potential, intra-city air pollution, Land Use Regression, machine learning, spatiotemporal

## Introduction

According to the 'National Human Activity Pattern Survey' (NHAPS), Americans spend about 87% of their time indoors (Klepeis et al., 2001). Heating, ventilating, and air conditioning (HVAC) systems are used to provide comfort conditions thermally and provide indoor air quality (IAQ) for the sake of the health and well-being of building occupants (Dutton et al., 2013) which accounts for 39% of the primary energy resources in the U.S. (*Annual Energy Review 2011*, 2012). Extensive efforts have been undertaken to develop methods and tools for measuring and understanding energy consumption and air quality within the built environment. However, limited studies have been done to integrate energy and air quality in a unified model to provide a comprehensive analysis considering outdoor conditions. This limitation can be more critical for those of naturally ventilated buildings in dense or polluted urban areas in which understanding the trade-offs and synergies between energy-saving goals and indoor air quality are essential. Integrating energy and air quality is a challenging task mainly due to the performance of existing tools in terms of accuracy and reliability, which consequently affects health and productivity of occupants in work environments.

Even though natural ventilation (NV) can potentially reduce energy loads during thermally suitable hours, improper design of NV systems results in certain acute and chronic diseases (Suh et al., 2000) due to bringing in the outdoor air pollutants. Particulate matters lesser than or equal to 2.5 μm in aerodynamic diameter ($PM_{2.5}$) as one of the major causes of morbidity and mortality has been reported to cause 4.2 to 8.9 million deaths per annum globally. However, 90% of people in developed countries spend their time indoors, 40% to 60% of total $PM_{2.5}$-related mortality in the U.S. still corresponds to outdoor-sourced PMs (Azimi & Stephens, 2018). According to the literature, the main body of existing studies relies on generalized urban scale outdoor air pollution data for indoor air quality assessments. This mainly stems from the lack of a robust approach to effectively connect the estimation of local outdoor-sourced PMs into the analysis of energy-air quality workflow.

As data is becoming more and more accessible along with an extensive global network of sensors, advances in AI technologies, and emerging smart cities, these advancements can provide a foundation for emerging, more accurate, faster, deeper, and actionable intelligence in design of the built environment. This approach can lead to performance-based and cost-efficient design decisions and reduce related human health risks. However, to date, limited work has been done to use this novel approach for integrating energy and air quality in a unified model. The current doctoral research aimed to develop a hybrid data-driven simulation framework for integrated Energy and outdoor-indoor Air Quality (iE-AQ) that combines novel data-driven approaches and AI techniques with conventional simulation workflows, and incorporates urban context to enable a more accurate and time-efficient model for assessment of energy-saving potentials of office buildings considering outdoor and indoor air quality variations at early stages of design.

## Materials and Methods

The current workflow for the integrated Energy-Air Quality (iE-AQ) modeling includes five main steps. The proposed framework is built upon a four-step model: (1) identifying spatiotemporal factors that explain outdoor air quality variation across urban areas, (2) localizing hourly outdoor air quality through Land Use Regression approach hybridized with Gradient Boosting Machine (LUGB), (3) aggregating localized outdoor $PM_{2.5}$ variation in urban neighborhood resolution, and (4) quantifying IAQ and energy-saving potentials for naturally ventilated buildings through coupling CONTAM and EnergyPlus simulation platforms. The proposed framework was tested on the prototype large-size office buildings provided by the U.S. Department of Energy (DOE) which was placed in all Chicago neighborhoods. Figure 1 illustrates the workflow of integrated energy and outdoor-indoor air quality execution for the proposed framework.

### IDENTIFYING THE KEY URBAN SPATIOTEMPORAL AND HUMAN-RELATED FACTORS

In the current research, six sets of urban factors were identified to estimate intra-urban air quality: (a) meteorological factors including hourly wind speed, wind direction, wind gust, temperature, relative humidity, atmospheric pressure, precipitation, solar radiation, and ultraviolet (UV) indicator obtained from local providers, (b) daytime and calendar features including weekdays, weekends, hour of day, and season, (c) land cover characteristics including intensity of open water, open space, barren lands, forests, grass, scrublands, and herbaceous based on their surface area, (d) mobility networks including intensity of highways, major and local roadways, railroads, number of bus and train stations, distance to the nearest airport and bus depot, as well as traffic profiles, (e) building characteristics including building typology, gross floor area, and year-built, and (f) census and housing factors including population, income per capita, household units, and household size.

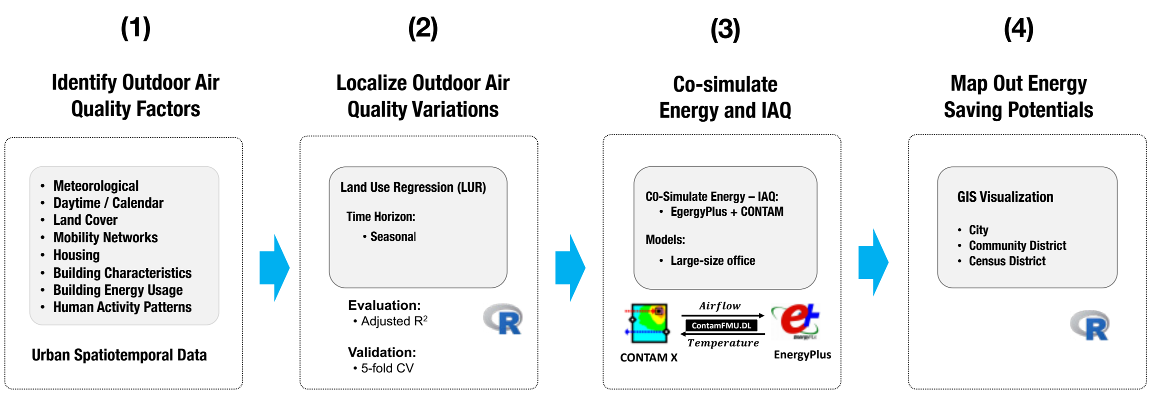


Figure 1: Workflow of integrated energy and air quality for indoor air quality and energy performance assessment in consecutive order.

### LOCALIZING OUTDOOR AIR POLLUTION ($PM_{2.5}$ CONCENTRATIONS)

Naturally ventilated buildings are often blamed for unexpected outdoor environmental conditions in providing "better" indoor environmental quality (IEQ) for occupants. The inconsistent wind speed and direction (Rong et al., 2015; Yang & Zhao, 2012), seasonal and severe weather variations (Shi et al., 2018), and outdoor air pollution are of critical limitations for performance evaluation of buildings rely on natural ventilation, of which the later limitation has been investigated less in the extant literature. The Land Use Regression (LUR) method is used to estimate the concentration of $PM_{2.5}$ for urban areas where no air monitoring stations exist to depict the outdoor air pollution profiles there. LUR is one of the robust methods in epidemiological health studies to estimate human exposure to urban air pollution (Hoek et al., 2008; Ryan & LeMasters, 2007) by which numerous spatiotemporal and auxiliary urban variables help localize air pollution profiles for entire urban areas. In this research 30 × 30 m grids are used for estimation of air pollution concentration across urban areas.

It should be noted that most LUR models are developed using statistical regression methods such as forward or backward stepwise regression approach in which the estimated air pollution values within a defined buffer radius are regressed on the values collected by a buffer-centroid air monitoring station in order to evaluate performance of LUR models. In this research, a LUR model based on Gradient Boosting Machine (GBM) (Greenwell et al., 2019; Friedman, n.d.) as one of the powerful machine learning (ML) techniques were utilized to increase the accuracy of intra-urban $PM_{2.5}$ estimation across urban areas. In the current research, outdoor $PM_{2.5}$ concentrations were localized based on seasonal aggregation of hourly data for weekdays and weekends, separately. In this research, the adjusted coefficient of determination (adj. $R^2$) was used to evaluate accuracy of the models.

Since LUR models employ multiple spatiotemporal variables, three criteria have been applied for feature selection: (a) the maximum variance inflation factor (VIF) (Allison, 1999), which reduces collinearity impacts, was set to 5 (VIF<10); (b) the maximum correlation coefficient (R) between variables in the same sub-category of variables and between different sub-category of variables were set to 0.6 and 0.95, respectively, captured from Dons et al. (2013); and c) buffer radii per variable was set between 50 m to 5000 m with 50 m intervals. It is worth noting that the lowest and highest buffer radii in LUR models are calculated for local road intensity and highway road intensity, respectively (Masiol et al., 2018). R programming software was used for executing the tasks for this step. Both LUR and LUGB models used cross-validation methods (Li et al., 2018) as an effective approach to avoid biased results. A 10-repeated 5-fold cross-validation (10 × 5CV) is used in this research by dividing data into train and test sets, including one-third of data as train set and two-thirds of data for validation.

### AGGREGATE AIR POLLUTION ($PM_{2.5}$) VARIATIONS ACROSS MULTI-SCALE URBAN BOUNDARIES

This step enables aggregating variations of $PM_{2.5}$ concentrations in multiple urban boundaries, herein in this article as neighborhood. In doing so, the spatial air pollution variation map which is provided in step 2, is averaged based on given spatial boundaries to find out the average value of that pollution within each boundary. This can also be implemented for finer intra-urban boundaries like census tract, census block, and building footprints. It should be noted that if the purpose of a design is only a single building in a specific location in a city, this step is not required and can be skipped. Thus, $PM_{2.5}$ variations for each of 30 × 30 m grids can be used in the next step (energy-air quality simulation) as inputs. R programming software was used for this step for computations.

### ENERGY-SAVING POTENTIAL AND IAQ ANALYSIS

The proposed framework couples CONTAM and EnergyPlus applications to solve combined interzonal airflow and heat transfer problems, effectively. With coupling these two applications through the Functional Mock-up Interface (FMI) platform, the hourly outdoor-sourced $PM_{2.5}$ concentrations and energy (cooling and ventilation) saving potentials during NV suitable hours are calculated for large-size office buildings. In this process, CONTAM computes inter-zonal airflows and passes it to EnergyPlus, then EnergyPlus calculates surface temperatures and gives it back to the CONTAM until the airflow and temperature for the model become converged. In this research, energy and IAQ are calculated at city and neighborhood (community) spatial resolutions, and then in the next step. Figure 1 step 4 illustrates the coupled CONTAM and EnergyPlus for simulation of IAQ and energy-saving potential for buildings. For setting natural ventilation characteristics in the EnergyPlus model, the inputs were taken from a recent study by Ben-David & Waring (2016), which uses DOE's prototype models. Further, the DOE's prototype model adjusted by the National Institute of Science and Technology (NIST) (Ng et al., 2019) was used to set interzonal parameters for the CONTAM model. Another study by Dols et al. (2016) elaborates coupling the two software through the FMI platform.

## Results and Discussions

Table 1 includes adjusted $R^2$ for hourly concentrations of $PM_{2.5}$ on a seasonal and annual basis through the LUR and LUGB methods. These results suggest 55.8%, 65.8%, 78.6%, and 60.4% improvements for explaining $PM_{2.5}$ concentrations during spring, summer, fall, and winter seasons, respectively, using the LUGB as the ML-based LUR approach. Further, it was found that $PM_{2.5}$ variations in the summer are lower than all other seasons. This is because of some natural factors like biogenic characteristics, which contribute to the $PM_{2.5}$ concentrations in hot seasons. For example, Lukacs et al. (2009) show that organosulfates as an organic factor account for 6% to 14% of $PM_{2.5}$ variation in summertime which confirms the complexity of $PM_{2.5}$ estimations for hot seasons because of unavailability of data. Further, the results considering differences between the two durations suggest that human activities during daytime more explain intra-urban $PM_{2.5}$ concentrations, which is confirmed by Dons et al. (2013).

After LUGB analysis which localized outdoor $PM_{2.5}$ variations across all urban areas using 30 × 30 m grids, the variations were then aggregated in neighborhood boundaries through averaging spatial variations of $PM_{2.5}$. In this research, DOE's large-size reference office building was used to calculate energy-saving potentials considering outdoor $PM_{2.5}$ concentrations. As explained earlier, CONTAM and EnergyPlus software were coupled to resolve mass balance and heat balance, simultaneously. Table 2 shows results of percentage of natural ventilation potentials in Chicago neighborhoods (min and max potentials) and city-scale resolutions. The results suggest that the energy saving potentials vary between 6.4% to 15.6% using localized $PM_{2.5}$ profiles aggregated into community districts.

| **Average Hourly Adjusted $R^2$** | | |
|---|---|---|
| | **Algorithm** | |
| | **LUR** | **LUGB** |
| Spring | 0.43 | 0.67 |
| Summer | 0.38 | 0.63 |
| Fall | 0.42 | 0.75 |
| Winter | 0.48 | 0.77 |
| Average | 0.43 | 0.71 |

Table 1: Average adjusted $R^2$ for seasonal hourly $PM_{2.5}$ concentrations based on LUR and LUGB algorithms.

| **Spatial Resolution** | **Energy-saving potential variation %** | |
|---|---|---|
| | **Min** | **Max** |
| Community district | 6.4% | 15.6% |
| City district | 9.18% | |

Table 2: Energy-saving potential variations in percentage for neighborhood and city scale resolutions based on LUGB method.

## Conclusion

The most significant outcome of this research is to propose a hybrid data-driven simulation framework for integrated energy and air quality modeling that incorporates climate change and health of occupants into the built environment design workflows through hybridizing AI-based data-driven techniques and physics-based tools. This enables a more accurate energy and air quality model, and at the same time, it increases the time and computational efficiency, which is considered as the main limitation of existing tools. Developing a LUGB allowed estimating outdoor air quality across urban zones where there is no air pollution information available. The proposed framework provided empirical evidence on energy-saving potentials for naturally ventilated buildings by considering outdoor air quality across urban neighborhoods. This model is applicable to all cities with similar available data. The proposed framework has the potential to aid designers, engineers, planners, and policy-makers with a better awareness of the existing profiles of outdoor air pollution at multi-scales. It also helps designers and decision-makers at the early stages of design to assess the feasibility of using natural ventilation responsive to energy-efficiency and health of building occupants in a time and cost-effective manner with a higher level of accuracy and resolution.

## References


Allison, P. D. (1999). *Multiple regression: A primer.* Pine Forge Press.

*Annual Energy Review 2011* (No. 0384). (2012). U.S. Department of Energy. https://www.eia.gov/totalenergy/data/annual/pdf/aer.pdf

Azimi, P., & Stephens, B. (2018). A framework for estimating the US mortality burden of fine particulate matter exposure attributable to indoor and outdoor microenvironments. *Journal of Exposure Science & Environmental Epidemiology.* https://doi.org/10.1038/s41370-018-0103-4

Ben-David, T., & Waring, M. S. (2016). Impact of natural versus mechanical ventilation on simulated indoor air quality and energy consumption in offices in fourteen U.S. cities. *Building and Environment, 104,* 320–336. https://doi.org/10.1016/j.buildenv.2016.05.007

Dols, W. S., Emmerich, S. J., & Polidoro, B. J. (2016). Coupling the multizone airflow and contaminant transport software CONTAM with EnergyPlus using co-simulation. *Building Simulation, 9*(4), 469–479. https://doi.org/10.1007/s12273-016-0279-2

Dons, E., Van Poppel, M., Kochan, B., Wets, G., & Int Panis, L. (2013). Modeling temporal and spatial variability of traffic-related air pollution: Hourly land use regression models for black carbon. *Atmospheric Environment, 74,* 237–246. https://doi.org/10.1016/j.atmosenv.2013.03.050

Dutton, S. M., Banks, D., Brunswick, S. L., & Fisk, W. J. (2013). Health and economic implications of natural ventilation in California offices. *Building and Environment, 67,* 34–45. https://doi.org/10.1016/j.buildenv.2013.05.002

Friedman, J. H. (n.d.). *GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE.* 44.

Greenwell, B., Boehmke, B., Cunningham, J., & GBM Developers. (2019). *Gbm: Generalized Boosted Regression Models. R package version 2.1.5. Https://CRAN.R-project.org/package=gbm* (Version 2.1.5) [Computer software]. https://CRAN.R-project.org/package=gbm

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., & Briggs, D. (2008). A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment, 42*(33), 7561–7578. https://doi.org/10.1016/j.atmosenv.2008.05.057

Klepeis, N. E., Nelson, W. C., Ott, W. R., Robinson, J. P., Tsang, A. M., Switzer, P., Behar, J. V., Hern, S. C., & Engelmann, W. H. (2001). The National Human Activity Pattern Survey (NHAPS): A resource for assessing exposure to environmental pollutants. *Journal of Exposure Science & Environmental Epidemiology, 11*(3), 231–252. https://doi.org/10.1038/sj.jea.7500165

Li, L., Chen, B., Zhang, Y., Zhao, Y., Xian, Y., Xu, G., Zhang, H., & Guo, L. (2018). Retrieval of Daily $PM_{2.5}$ Concentrations Using Nonlinear Methods: A Case Study of the Beijing–Tianjin–Hebei Region, China. *Remote Sensing, 10*(12), 2006. https://doi.org/10.3390/rs10122006

Lukacs, H., Gelencser, A., Hoffer, A., Kiss, G., Horvath, K., & Hartyani, Z. (2009). Quantitative assessment of organosulfates in size-segregated rural fine aerosol. *Atmos. Chem. Phys.,* 8.

Masiol, M., Zíková, N., Chalupa, D. C., Rich, D. Q., Ferro, A. R., & Hopke, P. K. (2018). Hourly land-use regression models based on low-cost PM monitor data. *Environmental Research, 167,* 7–14. https://doi.org/10.1016/j.envres.2018.06.052

Ng, L. C., Musser, A., Persily, A. K., & Emmerich, S. J. (2019). *Airflow and indoor air quality models of DOE prototype commercial buildings* (NIST TN 2072). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.TN.2072

Rong, L., Liu, D., Pedersen, E. F., & Zhang, G. (2015). The effect of wind speed and direction and surrounding maize on hybrid ventilation in a dairy cow building in Denmark. *Energy and Buildings, 86,* 25–34. https://doi.org/10.1016/j.enbuild.2014.10.016

Ryan, P. H., & LeMasters, G. K. (2007). A Review of Land-use Regression Models for Characterizing Intraurban Air Pollution Exposure. *Inhalation Toxicology, 19*(sup1), 127–133. https://doi.org/10.1080/08958370701495998

Shi, Z., Qian, H., Zheng, X., Lv, Z., Li, Y., Liu, L., & Nielsen, P. V. (2018). Seasonal variation of window opening behaviors in two naturally ventilated hospital wards. *Building and Environment, 130,* 85–93. https://doi.org/10.1016/j.buildenv.2017.12.019

Suh, H. H., Bahadori, T., Vallarino, J., & Spengler, J. D. (2000). Criteria Air Pollutants and Toxic Air Pollutants. *Environmental Health Perspectives, 108,* 9.

Yang, Y., & Zhao, Y. (2012). Prevailing Wind Direction Forecasting for Natural Ventilation Adjustment in Greenhouses Based on LE-SVM. *Energy Procedia, 16,* 252–258. https://doi.org/10.1016/j.egypro.2012.01.042