ARTICLE IN PRESS

Computers, Environment and Urban Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities

Constantine E. Kontokosta[a,b,*], Boyeong Hong[a,b], Nicholas E. Johnson[b,c], Daniel Starobin[d]

[a] Department of Civil and Urban Engineering, New York University, United States
[b] Center for Urban Science and Progress, New York University, United States
[c] University of Warwick, United Kingdom
[d] New York City Department of Sanitation, United States

ARTICLE INFO

Keywords:
Urban waste management
Municipal waste
Machine learning
Data analytics
GIS

ABSTRACT

Municipal solid waste management represents an increasingly significant environmental, fiscal, and social challenge for cities. Understanding patterns of municipal waste generation behavior at the household and building scales is a critical component of efficient collection routing and the design of incentives to encourage recycling and composting. However, high spatial resolution estimates of building refuse and recycling have been constrained by the lack of granular data for individual properties. This paper presents a new analytical approach, which combines machine learning and small area estimation techniques, to predict weekly and daily waste generation at the building scale. Using daily collection data from 609 New York City Department of Sanitation (DSNY) sub-sections over ten years, together with detailed data on individual building attributes, neighborhood socioeconomic characteristics, weather, and selected route-level collection data, we apply gradient boosting regression trees and neural network models to estimate daily and weekly refuse and recycling tonnages for each of the more than 750,000 residential properties in the City. Following cross-validation and a two-stage spatial validation, our results indicate that our method is capable of predicting building-level waste generation with a high degree of accuracy. Our methodology has the potential to support collection truck route optimization based on expected building-level waste generation rates, and to facilitate new equitable solid waste management policies to shift behavior and divert waste from landfills based on benchmarking and peer performance comparisons.

1. Introduction

Waste management is an increasingly complex quality-of-life issue for cities around the world, especially given the rapid growth of urban populations over the past two decades (World Health Organization Centre for Heath Development, 2010; Leao, Bishop, & Evans, 2004). Proper waste management is essential in order to provide sustainable, livable cities, as the collection and removal of waste impacts carbon emissions, traffic congestion, and air quality, as well as requiring significant operating expenditures (Adeyemi, Olorunfemi, & Adewoye, 2001; Esin & Cosgun, 2007). To improve waste management services and reduce the amount of waste sent to landfills, local governments are developing new methods to create efficient waste management systems and increase diversion rates through recycling and composting programs (MacDonald, 1996; Wang, Richardson, & Roddick, 1996; Bhargava & Tettelbach, 1997; Guerrero, Maas, & Hogland, 2013; Pappu, Saxena & Asolekar, 2007; Tam & Tam, 2006). New data streams, and the application of machine learning statistical methods, enable data-driven approaches to persistent problems in urban environmental management. Such data have proven to be a great resource for waste management planning, but they are typically collected at too coarse of an aggregation to fully optimize collection routing, and provide the empirical basis for policies that can shift, or "nudge", behavior through incentives or regulations based on performance metrics. The need for high resolution and targeted municipal solid waste management policy is crucial to minimize the future negative environmental impacts of urban waste.

Previous work has aimed to improve municipal waste management by using systems dynamics or data-driven modeling techniques to predict waste generation and identify factors that explain waste and recycling behavior. In particular, studies using temporal models with lagged waste generation data have performed well for prediction and forecasting, in large part due to the time series auto-correlation observable in waste generation rates at the regional or local level. Missing

* Corresponding author at: Department of Civil and Urban Engineering, New York University, United States.
E-mail address: ckontokosta@nyu.edu (C.E. Kontokosta).

https://doi.org/10.1016/j.compenvurbsys.2018.03.004
Received 27 November 2017; Received in revised form 6 March 2018; Accepted 7 March 2018
0198-9715/

Please cite this article as: Kontokosta, C., Computers, Environment and Urban Systems (2018), https://doi.org/10.1016/j.compenvurbsys.2018.03.004

from the literature, however, are attempts to predict waste generation for individual buildings in a large-scale municipality. Part of the challenge emerges from data constraints, as few sanitation agencies collect and make available granular waste collection data. Furthermore, small-area estimation problems can confound attempts to accurately down-scale predictions from the city or district to individual buildings.

Given this context, this research attempts to inform municipal waste management operations by developing high spatial and temporal resolution estimates of waste and recycling generation rates for individual residential buildings. Using an extensive and granular waste collection dataset from the New York City Department of Sanitation (DSNY), coupled with detailed land use, demographic, socioeconomic, and weather variables, we are able to overcome previous data limitations to build a socio-spatial machine learning model to predict building-specific waste generation rates based on derived building population estimates for more than 750,000 residential properties in New York City. We validate our prediction using a sample of individual truck route collection data for specific days and locations that are representative of the City's land use types and densities. The results of the model and validation indicate that our method performs well at estimating building-level waste generation. Our findings also provide a comprehensive understanding of the socioeconomic and land use drivers of municipal waste generation in New York City and can enable more efficient, and equitable, waste management practices, particularly through route optimization and peer comparison benchmarking programs.

We begin by presenting a literature review of previous studies that attempt to predict waste generation using machine learning techniques, as well as applications of small unit estimation to determine building and household population size. Section 3 includes a description of our data and machine learning methods. Results are presented in Section 4, followed by a discussion of the findings and their implications for urban waste management and data-driven environmental policy.

2. Background

Numerous studies have used available waste data to identify significant factors that influence refuse and recycling rates, and to develop statistical models to forecast waste generation. Previous work has demonstrated that a broad range of factors drive waste generation depending on the particular study area, although most focus on macro-scale analyses of regions or metropolitan areas. Keser, Duzgun, and Aksoy (2012), for example, found that regional characteristics such as the unemployment rate, the asphalt-to-paved road ratio, temperature, higher education ratio, and agricultural production values have a significant influence on waste generation in Turkey. In Xiamen, China, Zhang et al. (2015) identified population, land use, and building coverage as important factors driving waste generation. Oribe-Garcia, Kamara-Esteban, Martin, Macarulla-Arenaza, and Alonso-Vicario (2015b) support and extend these studies by suggesting that urban morphology, tourism activity, educational level, economic status, and resources of the population impact aggregate residential waste generation. Denafas et al. (2014) estimate seasonal variations of waste generation in Eastern European cities by using time series forecasting models. Their results suggest that geographical latitude is a major factor in the seasonal pattern of waste generation across cities primarily due to differences in local weather.

A variety of models have been developed to predict waste generation, typically at the national or city scale. Karadimas and Loumos (2008) used the ant colony system algorithm to predict waste generation based on spatially-dependent characteristics such as the location of waste bins, the road network topology, and population density. Other researchers have built predictive models to forecast waste generation using temporal features such as seasonality and historical trends. For instance, Rimaityte, Ruzgas, Denafas, Racys, and Martuzevicius (2012) found that an Autoregressive and Integrated Moving Average (ARIMA) model combined with seasonal exponential smoothing is an effective method to predict weekly waste generation. Modern machine learning methods, such as neural networks, have also been used in temporal models to predict waste generation. Zade and Noori (2008) for example, used a feed-forward artificial neural network (ANN) to forecast weekly waste generation in the tourist city of Mashhad, Iran. They found that ANNs perform well when predicting waste generation at low spatial resolutions, and introducing time lag features into an ANN model can address serial correlation in the time series data. Antanasijevic, Pocajt, Popovic, Redzic, and Ristic (2013) also used ANNs to predict solid waste generation at the national level in Bulgaria and Serbia.

Combinations of spatial and temporal models have also been used to improve prediction results. For instance, Dyson and Chang (2005) developed a systems dynamics model using population, median household income, household size, and historical waste generation data to predict waste generation in San Antonio, Texas. In a particularly relevant study, Johnson et al. (2017) developed a spatiotemporal model using a gradient boosting regression tree algorithm and features such as weather, urban morphology, and socioeconomic and demographic information to predict weekly waste generation for 232 administrative sections (geographic divisions) in New York City. The results demonstrate high predictive accuracy for waste generation using historical waste generation rates.

These studies, however, are often limited by spatially-aggregated data that do not allow for analysis or prediction of waste generation patterns across small areal units. Data are typically aggregated into larger geographical units due to privacy and confidentiality issues, or the lack of data collection at the truck, household, or building scale. Previous attempts at building or household population estimation tend to rely on surveys (Ojeda-Benítez, Armijo-de Vega, & Marquez-Montenegro, 2008; Thanh, Matsui, & Fujiwara, 2010) or on down-sampling from predictions at larger geographies, such as the city or province (Oribe-Garcia, Kamara-Esteban, Martin, Macarulla-Arenaza, & Alonso-Vicario, 2015a; Purcell & Magette, 2009). These studies are either difficult to generalize, given their small samples sizes and limited historical data, or do not account for significant variations in the occupancy of specific buildings that can be obscured by relying solely on population surveys, such as census data. This is a significant gap in the literature, as building-level waste generation data can provide important insights for collection route optimization, household waste benchmarking programs, and more equitable waste reduction incentive and behavior change initiatives built on information transparency.

3. Materials and methods

This study aims to predict weekly and daily municipal waste generation from residential properties at the building level using a data mining and machine learning approach. We first develop a predictive model by comparing the performance of gradient boosting regression tree (GBRT) and Neural Network (NN) machine learning algorithms to estimate weekly waste generation for each of the 609 *DSNY sub-sections*, which are collection areas (also known as "frequencies") within the 232 DSNY sections. We then estimate individual building populations for all residential properties in NYC by implementing small area estimation methods that combine census population data with specific building characteristics including type, size, and density. Weekly generation at the building-level is then calculated by multiplying the predicted per capita weekly waste generation for each DSNY sub-section with the estimated building population of a given building located within that DSNY sub-section. This approach accounts for inter-sub-section variations in waste generation behavior, driven by such variables as demographics or socioeconomic characteristics, and building-specific factors, such as the number of residential units and their relative size. Fig. 1 presents a flow chart of our methodology.

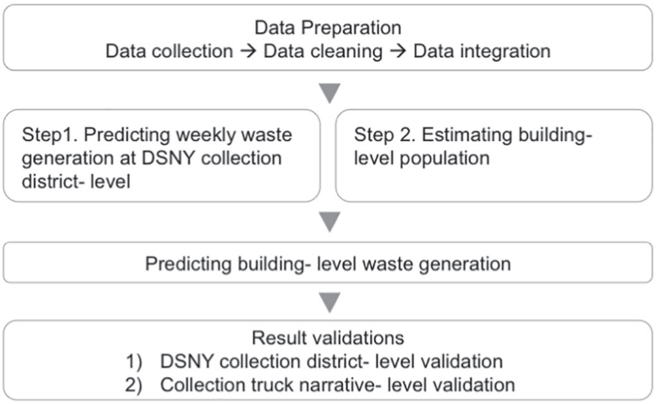

Fig. 1. Methodological approach to building-scale waste generation estimates.

3.1. Data

In 2016, 3.7 million tons of municipal solid waste was collected from residential and institutional buildings in New York City, accounting for 25% of all waste generated (New York City Department of Sanitation, 2017b). According to DSNY, each resident, on average, generates 11.14 pounds of refuse and 2.46 pounds of recycling material per week (The Council of the City of New York, 2017). Currently, over 80% of waste collected by DSNY is disposed of in landfills at a cost of nearly $400 million dollars a year (Johnson et al., 2017). To reduce the cost and environmental impact of landfilling, New York City's OneNYC plan has established a goal of sending zero waste to landfills by 2030. To meet this target, DSNY has launched new programs and services to encourage and incentivize residents to reduce overall refuse generation and increase recycling rates. One notable initiative is the pilot organics program to collect food scraps and yard trimmings to be used for renewable energy (New York City Department of Sanitation, 2017b).

DSNY provides regular curbside waste collection for residential buildings, public schools, city-owned buildings and certain nonprofit organizations, with waste categorized into refuse, paper and metal/glass/plastic (MGP) streams (New York City Department of Sanitation, 2017a). In all, DSNY is responsible for collection from 755,594 of the 858,370 properties (88%) in New York City. DSNY operates curbside waste collection by using weight-based truck collection (tonnage), and each truck is designed to accommodate 12.5 tons of refuse, 11.5 tons of paper and 10.0 tons of metal, glass, and plastic (New York City Mayor's Office of Environmental Coordination, 2014). DSNY divides the city into 232 sections that are further subdivided into 609 spatial areas based on collection schedules, called sub-sections. Fig. 2 shows this waste collection geography. DSNY tracks daily truck-route waste tonnage for each type of waste, and has provided historical records of sub-section waste collection data from 2013.

Table 1 describes the primary data used in this work, which includes waste data supplied by DSNY with daily collection tonnages for three waste streams from January 2013 to November 2016 for each of the 609 DSNY sub-sections. Since individual geographies have different collection schedules, such as bi-weekly or tri-weekly (Johnson et al., 2017), we aggregate the daily collection data to the week in order to compare different DSNY sub-sections across the city based on a common temporal scale. Waste data are integrated with the New York City Department of City Planning Primary Land Use and Tax Lot Output (PLUTO) dataset, which describes individual tax lots (properties) and includes the building class, land use, owner type, building area, the number of units, and the construction year, among other features. Though we consider the tax lot (based on the borough-block-lot or BBL identifier) as a proxy for an individual building, a tax lot can have more than one building. Across NYC, the number of lots with multiple buildings is 223,841, which is 26.1% of all properties.

Population and socioeconomic characteristics were obtained from

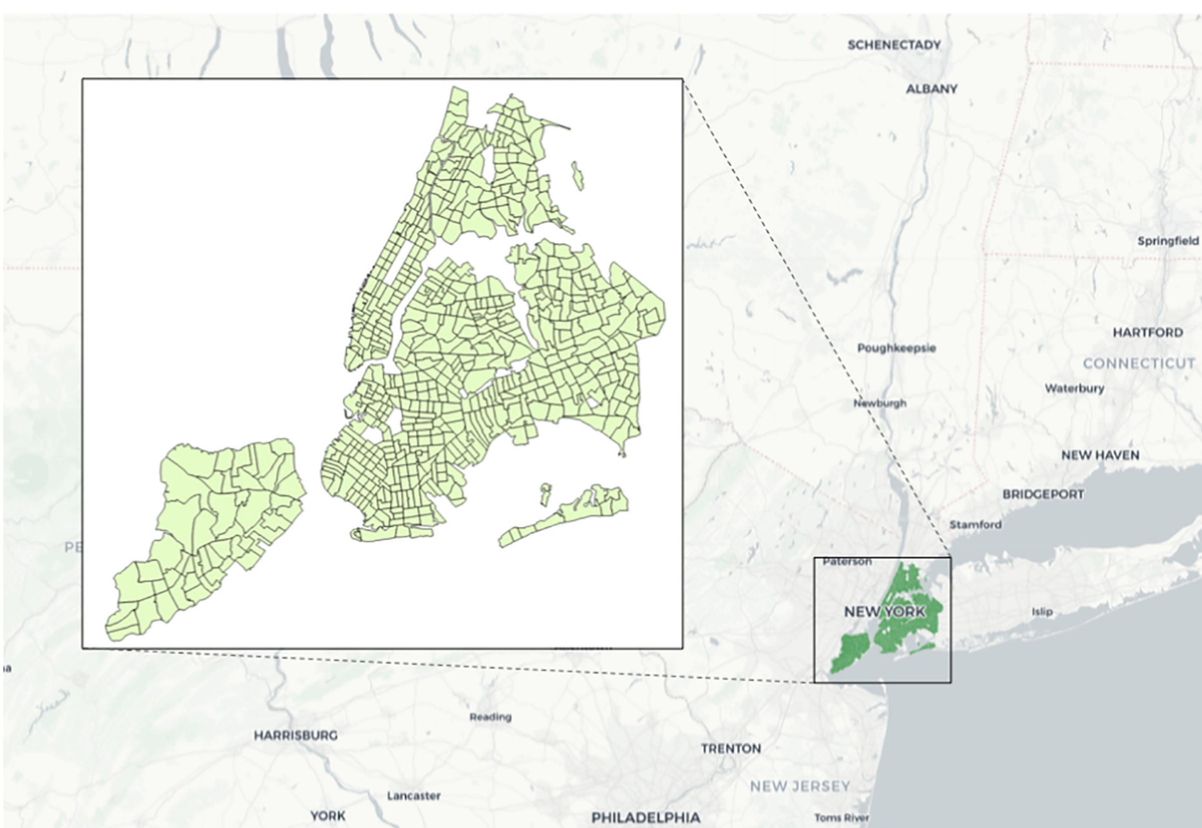

Fig. 2. Study area - New York City divided into 609 DSNY sub-sections (administrative collection districts).

Table 1
Description of data sources.

Category	Data	Description	Year	Source
Waste	Waste collection data	Waste collection tonnage data (for 232 sections and 609 sub-sections) for residential and institutional properties	2013–2016	New York City Department of Sanitation
Urban form	MapPLUTO	New York City Tax Lot attributes including the number of units, residential area, ownership, building class, land use, year built, etc.	2016	New York City Department of City Planning
Demographic and socioeconomic	American Community Survey (ACS)	Demographic and socioeconomic data such as population, gender, age, income level, race, and education attainment	2015 5-year estimate	U.S. Department of Commerce
	DOE school enrollment data	Number of current students of each public school	2016	New York City Department of Education
	Local Law 84 data	Building data including the number of municipal employees and the number of bedrooms of residential buildings over 50,000 ft^2	2016	Mayor's Office of Sustainability
Weather	Weather data	Weatherdata including temperature, precipitation, snow, wind speed etc.	2013–2016	Weather Underground
Holiday	Holiday data	Dates of U.S. holidays		Python package

the 2015 5-year estimate American Community Survey (ACS). Data from ACS include information on households at the Census Block Group (CBG) level, of which there are 6100 in NYC. We include population by gender, age, race, primary language, employment status, educational attainment, household type, median income, median rent, and vacancy rate data. For population estimates in non-residential (municipal and public school) buildings serviced by DSNY, the number of current students for each public school was obtained from the New York City Department of Education (DOE), and the number of employees in municipal buildings was provided by the NYC Mayor's Office of Sustainability (MOS), as collected through Local Law 84 (Kontokosta, 2015). Historical weather information and U.S. national holidays were used to control for the seasonal variations in solid waste generation. Weather data include temperature, precipitation, wind speed, and snow events. These data are resampled to a weekly moving average by using the mean or sum calculation in order to match the temporal frequency of our waste generation data. Similarly, the number of U.S. holidays are computed for each week.

3.2. DSNY sub-section waste prediction model

A machine learning model is used to predict each waste stream, as well as total waste generated, at the DSNY sub-section scale. Table 2 provides a complete list of the features used to predict weekly waste

Table 2
Description of extracted predictors.

Category	Predictor	Variable description
Urban form	City-owned building	Percentage of city-owned building
	Commercial space	Percentage of commercial area
	Retail space	Percentage of retail area
	Residential space	Percentage of residential area
	Vacancy	Vacancy rate
	Building age	Median year of construction
	Residential units	Number of residential units
	Public housing	Percentage of public housing area
	One family housing	Percentage of one family-housing area
	Building density	Median as-built FAR (the ratio of total gross floor area to the area of its zoning lot)
Demographic and socioeconomic	White	Percentage of White population
	Black	Percentage of Black population
	Asian	Percentage of Asian population
	Racial diversity	The ratio of different racial groups in the geographical unit
	Employment	Employed rate
	Education level	Percentage of population with at least bachelor's degree
	Elderly population	Percentage of population over 65
	Female	Percentage female population
	Households with children	Percentage of households with children under 6
	Households living alone	Percentage of households living alone
	Limited English speakers	Percentage of limited English speakers
	Rent	Median gross rent
	Income	Median household income
	School enrollment	Number of students enrolled in public schools
	City employees	Number of city employees in municipal buildings
Seasonality	Temperature	Average temperature
	Precipitation	Total precipitation
	Snow	Total snowfall amount
	Wind speed	Average wind speed
	Weather event	Number of severe weather events
	Holidays	Number of U.S. holidays

Table 3
Building typology (NYC Department of Building).

Building typology	Definition
One-family house	A building designed for and occupied exclusively by one family
Two-family house	A building designed for and occupied exclusively by two families
Walk-up apartment	Multiple dwelling building without elevator
Elevator apartment	Multiple dwelling building with elevator
Condominium apartment	Multiple dwelling with separate ownership
New York City Housing Authority (NYCHA) public housing	Affordable housing (multiple dwelling) for low income residents supported by NYCHA

generation (dependent variable) in tons for each of the 609 DSNY sub-sections. The model includes a total of 31 features and are trained on data from 2013 to 2015 and tested on 2016 data. Features were selected based on domain knowledge and previous literature, and feature importance based on the model results. We initially test GBRT and NN models with standardized data to evaluate model performance. GBRT has been shown, in general, to out-perform other ML methods, such as support vector regression and Random Forest models (Friedman, 2001; Schapire, 2003), in similar applications, thus we do not test these algorithms here. GBRT limits over-fitting through hyperparameter tuning, and is able to account for complex, non-linear relationships between variables, an improvement over simple linear approaches. Linear models are also very susceptible to outliers in the data, as well as collinearity within features, unlike GBRT. GBRT is an advanced iteration of a typical decision tree model, with the features recursively split and re-weighted based on the errors of the previous trees, and the goal being the reduction of the residual sum of squares (Johnson et al., 2017; Ye, Chow, Che, & Zheng, 2009).

We compare the initial performance of the GBRT and NN models based on R-squared and Root Mean Squared Error (RMSE) values; the GBRT model yields 0.87 R-squared and 0.034 RMSE, and the NN model yields 0.77 R-squared and 0.050 RMSE. We therefore select GBRT for model development and further parameter tuning. Model parameters are tuned using a grid search in order to maximize model performance. We focus our tuning on the number of trees, tree depth, and learning rate, which control the model's structure and complexity. The final model takes the number of trees = 200, tree depth = 6, and learning rate = 0.1. Also, an iterative cross-validation method is implemented to ensure the robustness of the model. After tuning model hyperparameters, we compute estimates of weekly waste generation across three waste streams (refuse; paper; and metal, glass and plastic) and total waste generation (sum of all waste streams) for each DSNY sub-section in 2016.

3.3. Building population estimation

The second step in our methodology is to estimate the number of occupants per building for residential properties serviced by DSNY. For public school buildings and municipal office buildings, we used existing data from the NYC DOE and NYC MOS Local Law 84, respectively. Reliable estimation of residential building population, however, is a persistent problem in urban management and planning (Lwin & Murayama, 2009). Currently, the most granular level of population data, with reasonable margins of error, is U.S. Census ACS data aggregated to the Census Block Group (CBG). Since no building-level population data are available, we develop an approach to estimate individual building populations using existing data from various spatial scales, from building-specific physical and land use characteristics to CBG population estimates. This approach is designed to be generalizable beyond the waste prediction application presented here.

To estimate residential building populations, we first match all buildings using their spatial identifier (BBL) in the PLUTO dataset to their respective CBG. This results in each BBL being assigned a specific CBG identifier. The population of each building is then estimated by multiplying the computed occupant density for each building type in a CBG by the total residential floor area of the given building. Table 3 provides a list of building types found in the PLUTO dataset. The final calculation takes the form

$$\widehat{P} = A \times \widehat{D}_{nm} \tag{1}$$

where $\widehat{P}$ is the estimated building population, A is the total residential floor area of the building and $\widehat{D}$ is the calculated occupant density for building type n in CBG m. Because the actual occupant density is not known for each building, we compute the occupant density across building types at the CBG-level using an Ordinary Least Squares (OLS) regression model. This method is designed to account for significant differences in occupant density across different residential building types (e.g. two-family home to high-rise apartment building), and for socioeconomic and demographic variations across neighborhoods. For example, based on the total floor area per occupant, walk-up apartment buildings are found to have relatively higher population densities when compared to single-family homes. In order to capture this variation, we perform a multivariate linear regression that controls for building type. We build on a similar dasymetric method implemented by Langford, Maguire and Unwin (1991) and Langford (2013), by adding individual building-specific characteristics as ancillary data. The regression model takes the form

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{2}$$

where y is the total population of the CBG and x is the total residential floor area of the CBG for each housing type n. The coefficients β_n are computed and the ratios of the coefficients are used to weight the corresponding occupant density for each housing type in a given CBG. The total CBG population is therefore allocated to different building types based on 1) the total area of that building type in the CBG and 2) the ratio of the city-wide coefficients. This method builds on previous small area population estimates by including building-specific type and size characteristics, while accounting for socio-spatial variations in occupant density based on demographic characteristics of a given neighborhood.

3.4. Model validation

Since there are no collected or "ground truth" data available for waste generation of individual buildings, it is impossible to evaluate the model results against actual building-level data. However, we introduce two alternative approaches to validate our model results using DSNY sub-section and *route narrative* (individual collection truck) waste data. The first validation method aggregates our building-level waste predictions to the DSNY sub-section. The sum of the building-level estimates is compared against the actual waste collected for the DSNY sub-section for the respective time period. Although there are limitations in this approach, such as the possibility that under- and over-estimates at the building level can cancel out when aggregated, it does provide bounds for the error of our prediction.

The second validation method uses individual truck collection data to provide a more robust and application-oriented evaluation of the proposed model. We first assign contiguous tax lots (buildings) to

Fig. 3. Building-to-truck route matching (DSNY sub-section 1021A – Monday case).

Table 4
DSNY sub-section prediction model results.

	Total waste	Refuse	MGP	Paper
R-squared	0.87	0.87	0.73	0.78
Samples with less than 20% error	85.3%	82.3%	78.9%	79.8%

individual truck routes. Unfortunately, there is no information available on the specific set-out point for individual buildings, thus adding uncertainty to the assignment of buildings to specific trucks. For tax lots with property lines adjacent to two truck routes, we assume that the larger property line dimension is the primary set-out point for waste collection. In this way, each DSNY sub-section is divided into multiple truck routes for each collection day. We then extrapolate daily waste generation by applying the ratio of average waste generation for each day of the week, obtained from the actual sub-section-level collection data, to the predicted weekly waste generation at the building-level. The per truck waste generation is then computed as the sum of the predicted waste generation for each building associated with a given route for a given day. This total is then compared to the actual truck collection tonnage for that day. An illustration of this building-to-truck route matching is shown in Fig. 3.

The truck-level validation was performed for two DSNY sub-sections, one in Manhattan (1021A) and one in Brooklyn (3114C). These two routes were selected for data completeness (with respect to individual routes) and the heterogeneity of building densities within the two areas. The Manhattan route consists of mixed-use, high-density buildings, while the Brooklyn route consists primarily of low-density residential buildings. Furthermore, these two areas are representative of DSNY service areas across the city. We do not attempt additional route validations because data from DSNY are limited and received in text (PDF) form. Individual truck routes, and the buildings adjacent to those routes, had to be digitized using an algorithm developed to extract text files and spatially join each respective street segment and route location. The reliability and consistency of the route data provided by DSNY limited a more expansive truck-level validation.

4. Results

4.1. Predictive model performance and feature importance

Table 4 shows a summary of the performance metrics for the four predictive models for total waste, refuse, MGP recycling, and paper recycling for DSNYsub-sections in 2016. All models perform well, with high R-squared values (mean of 0.81) and a high proportion of samples with less than 20% mean absolute error (MAE). Figs. 4–7 show our predicted weekly waste generation for each DSNY sub-section against the actual value from the DSNY dataset.

The total waste generation model performed the best with an out-of-sample R-squared value of 0.87 and 85.3% of sub-sections with a MAE

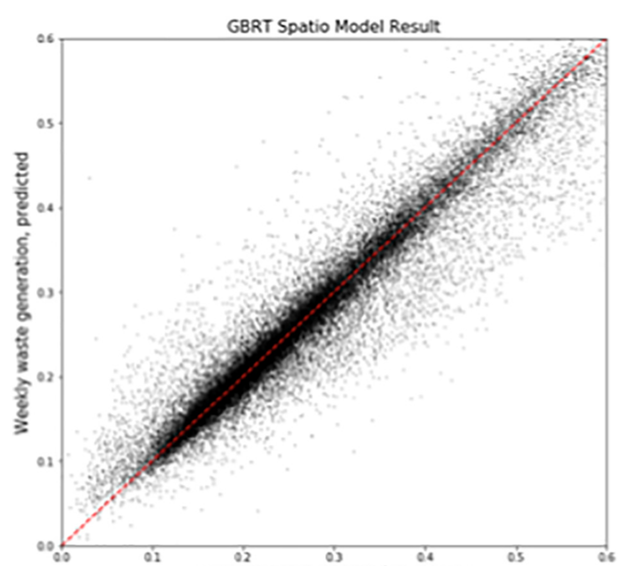

Fig. 4. Scatter plot of predicted v. actual values: Total waste.

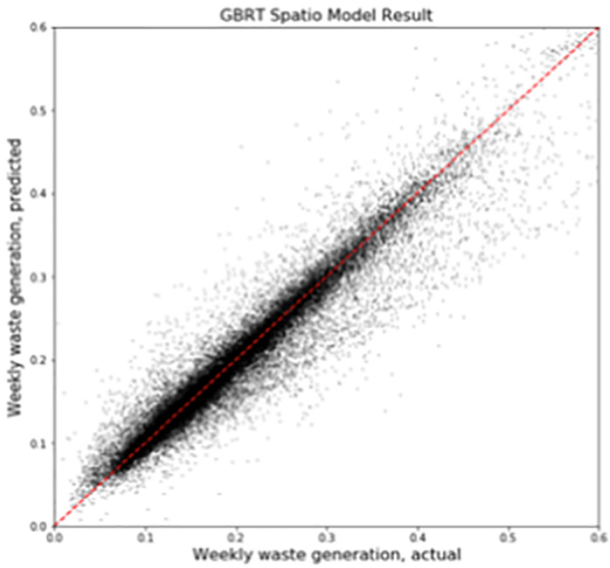

Fig. 5. Scatter plot of predicted v. actual values: Refuse.

of less than 20%. The total waste and refuse models showed similar performance, which is in large part due to the fact that refuse comprises approximately 80% of the total solid waste stream. The R-squared for the refuse model is 0.87 with 82.3% of samples less than 20% MAE, which is slightly lower than the total waste prediction model in part because of variance in recycling rates across the City. As expected, these two models perform better than the two recycling stream models. The MGP and paper recycling models showed lower prediction accuracy with R-squared values of 0.73 and 0.78, respectively. The reduced performance of the recycling models reflects the significant spatial variations in recycling behavior. For example, Fig. 9 shows a discernible spatial pattern where the highest recycling rates are located in parts of Manhattan, Downtown Brooklyn, and Staten Island, while the lowest recycling rates correspond to some of the poorest communities in the city. These spatial patterns reflect the high correlation between

Fig. 6. Scatter plot of predicted v. actual values: MGP.

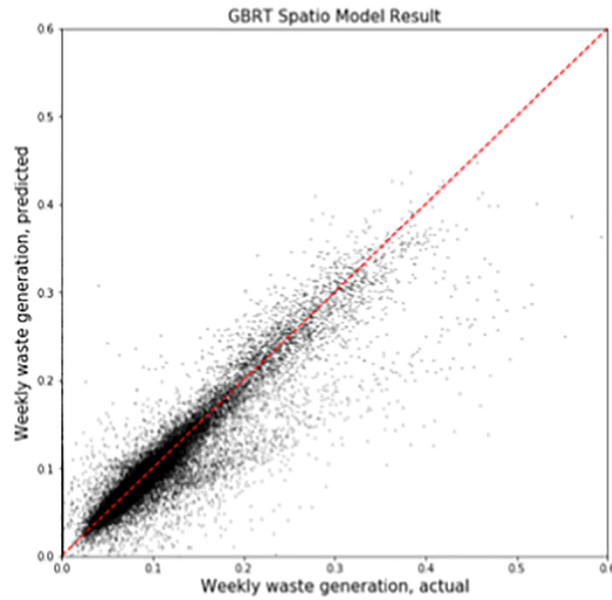

Fig. 7. Scatter plot of predicted v. actual values: Paper.

recycling rates and socioeconomic attributes, including income and educational attainment. Refuse generation per capita, however, follows a more normally-distributed pattern as seen in Fig. 8.

One interesting finding of the predictive model, which reinforces previous results, is that weather variables (temperature, precipitation, wind speed, and snow) are ranked highly as important features for all prediction models. Certainly, weather influences waste collection both in terms of seasonality in waste generation, and in its effect on collection totals on days with severe weather conditions. Population, residential area, household characteristics, and education level are also shown to be important features in the weekly waste generation prediction model.

4.2. Building population estimation

Table 5 shows the OLS regression results for computing the relative occupant densities by housing type. The ratio of the residential density of the six different housing types - (1) one-family house, (2) two-family house, (3) walk-up apartment, (4) elevator apartment, (5) condominium apartment, and (6) NYCHA public housing - is found to be 19:32:35:10:12:19. The model is able to explain much of the variance in occupant density (R-squared value is 0.84) and all coefficients are statistically significant at the 95% confidence level. Elevator apartments have relatively lower occupant densities, while walk-up apartments tend to have high occupant densities. By using the population of each CBG, the ratio of occupant densities of the six different housing types, and the sum of residential gross floor areas of each housing type, we are able to estimate individual building populations. An example is shown in Table 6.

In order to validate our building population estimates, we use data from NYC Local Law 84, which is the building energy benchmarking ordinance covering all buildings larger than 50,000 ft^2 (Kontokosta & Tull, 2017; Reina & Kontokosta, 2017). In addition to reporting annual energy use, residential building owners are required to provide the total number of bedrooms in the building, and we use this bedroom data to provide bounds to our building population estimates. Although only large multi-family buildings are included in the validation data, as

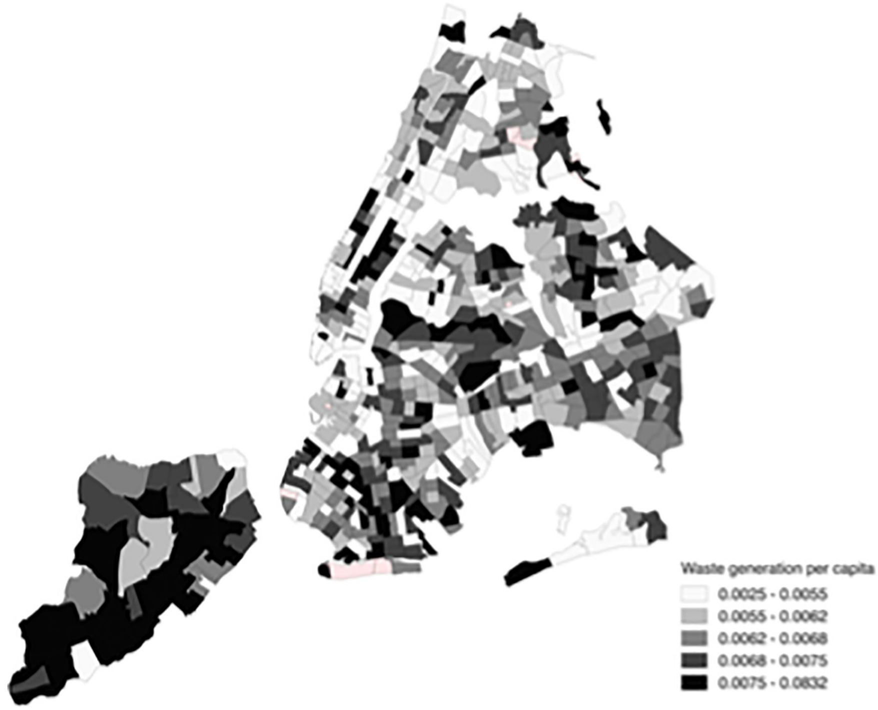

Fig. 8. Weekly average waste generation per capita for 609 DSNY sub-sections (2013–2016, tons).

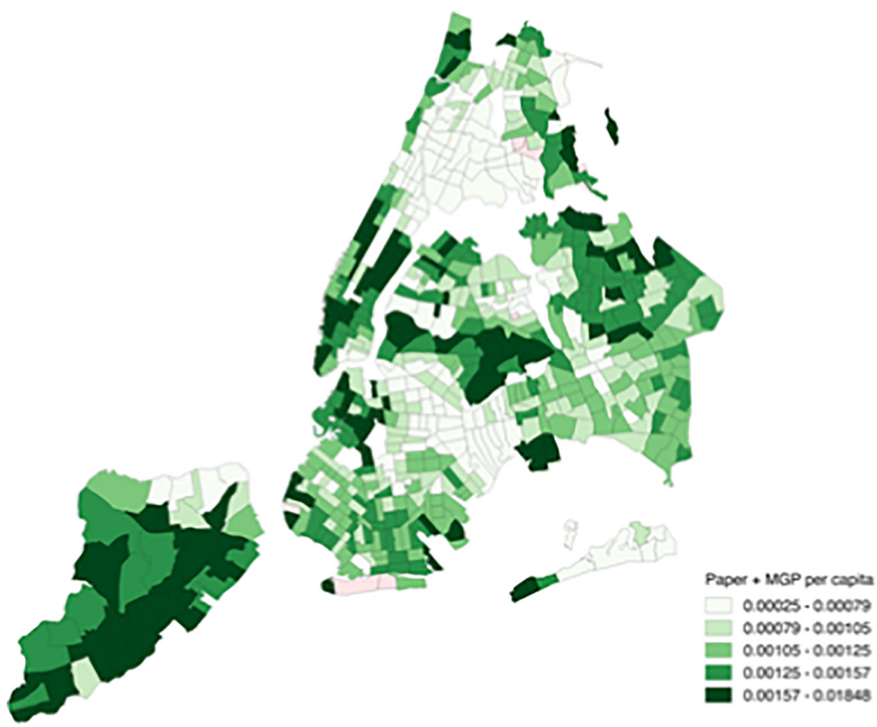

Fig. 9. Weekly average recycling per capita for 609 DSNY sub-sections (2013–2016, tons).

Table 5
OLS results: ratio of corresponding occupant density based on housing type.

Residential building typology	Coefficient
One-family house	0.0019*
Two-family house	0.0032*
Walk-up apartment	0.0035*
Elevator apartment	0.0010*
Condominium apartment	0.0012*
New York City Housing Authority (NYCHA) public housing	0.0019*

* p-value < 0.05.

Table 7 shows, the estimated population demonstrates a reasonable relationship to the reported number of bedrooms.

4.3. Predicted solid waste generation for individual buildings

Building waste generation is calculated by multiplying the predicted weekly waste generation per capita for a given CBG by the estimated occupant population for a given building. Table 8 shows an example output for the week of October 24th, 2016. The first column is the BBL identifier for an individual property, the second column is the DSNY collection schedule number, the third column is the estimated number

Table 6
Example output of population estimation.

BBL	Bldg class	Res area	D_A	D_B	D_C	D_D	D_R	D_P	POP_{est}
3070400033	A	2064	0.0020	0.0034	0.0036	0.0010	0.0013	0.0020	5
3041510051	B	1608	0.0024	0.0040	0.0043	0.0012	0.0015	0.0024	7
3041520024	D	438,356	0.0018	0.0030	0.0034	0.0009	0.0012	0.0024	424

Table 7
Example outcome: Validation of building population estimation output - number of bedrooms per building vs. estimated population.

BBL	Bldg class	Number of bedrooms (LL84 data)	Estimated population
4020860040	D	500	460
3056460036	D	91	82
3011890007	D	139	158
3047750041	D	66	64
1015320001	D	260	353
3050620014	D	134	138
3007290072	D	82	82

of building occupants, the fourth column is the predicted weekly waste generation per capita for the DSNY sub-section, and the last column shows the predicted value of the total weekly waste generation (in tons) for that building.

Fig. 10 show choropleth maps of the prediction results across the four different waste streams for the week of October 24th, 2016 for DSNY sub-section 1021A, located in the SoHo neighborhood of Manhattan. Yellow to green colored parcels are serviced by DSNY, and the darker color represents a larger volume of weekly waste generation for those properties.

Fig. 11 shows the maps of the average errors between the model results and actual data from October to November 2016. The red subsections indicate where the model predicts greater refuse than actual, and blue regions show an under-estimation. A comparison of the

Table 8
Example output of building-level weekly total waste generation prediction.

BBL	DSNY sub-section	Bldg occupants	Predicted weekly waste generation per capita (tons)	Predicted total weekly waste generation (tons)
1021790466	1122B	144	0.004646	0.669
1021680061	1123A	107	0.004662	0.499
1021740182	1123A	77	0.004662	0.419
1021540088	1123B	8	0.005442	0.044
1021700340	1123A	125	0.004662	0.583

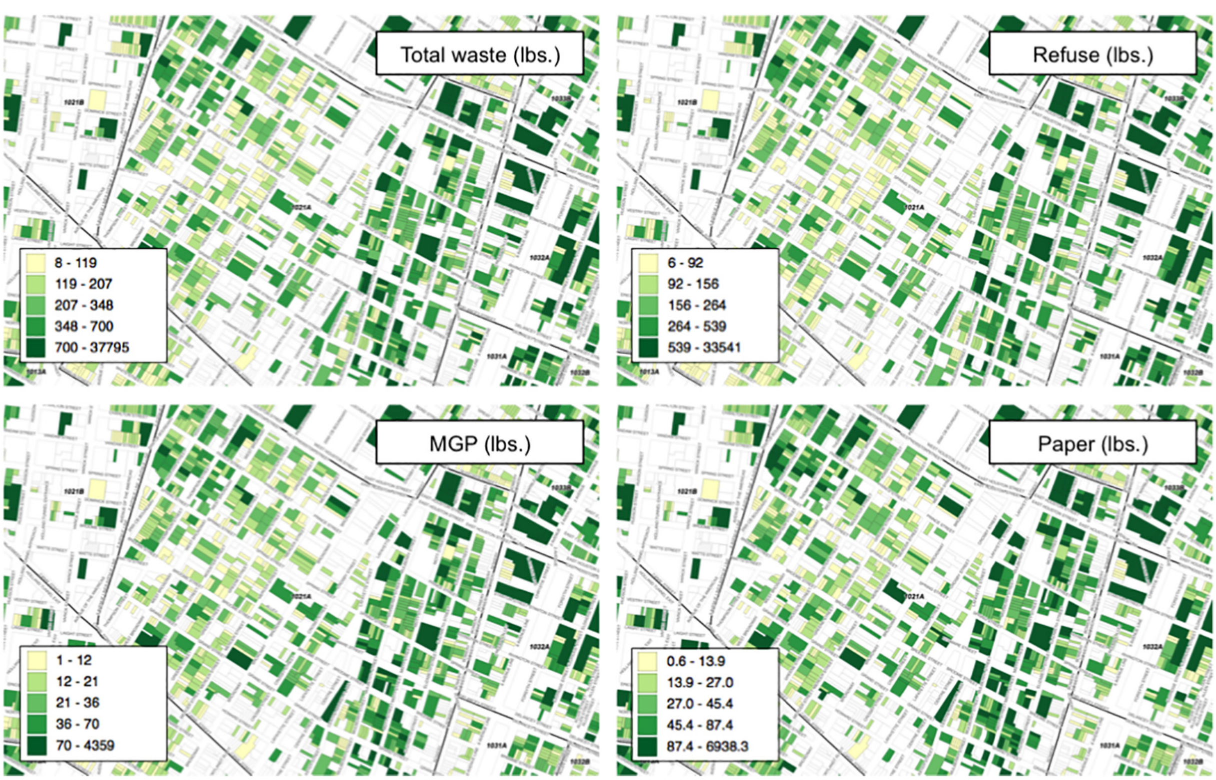

Fig. 10. Weekly waste generation prediction results (DSNY sub-section 1021A case: Monday, October 24th, 2016).

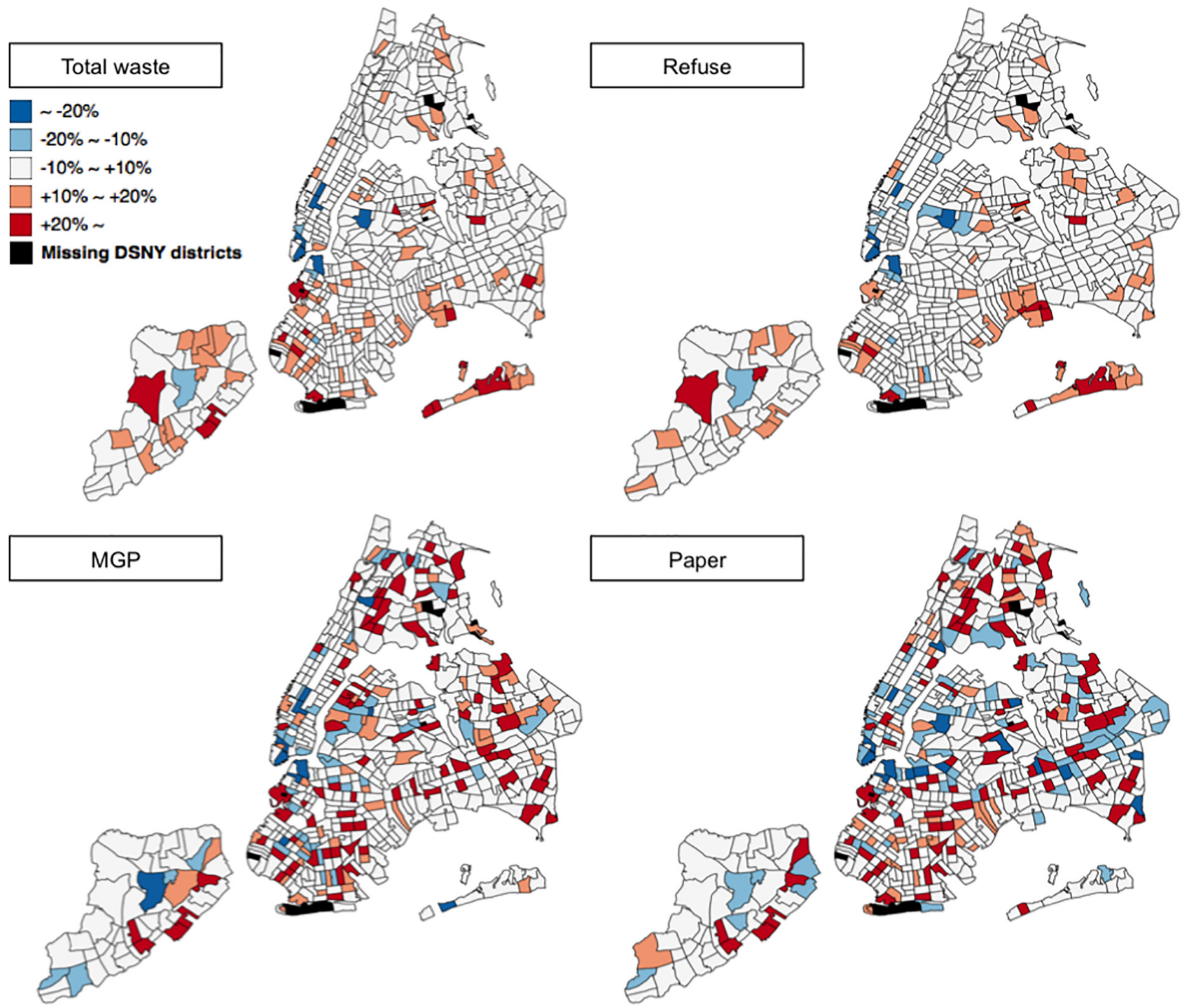

Fig. 11. DSNY sub-section level weekly waste generation validation - the maps of the average errors between the model results and actual data from October to November 2016.

visualizations in Fig. 11 shows that there, as expected, are higher error rates for recycling material generation than refuse generation alone. Overall, we find that the projected total waste and refuse generation are within 10% of the actual waste generation in 83% of the sub-sections.

The truck route validation using the two DSNY sub-sections described above indicates that the proposed model performs well in terms of absolute error. Fig. 12 visualizes the different truck routes on a Monday in DSNY sub-section 1021A (Manhattan) for refuse collection. Colored polylines represent individual trucks, and the line width represents the relative amount of refuse generated for each truck. Tables 9 and 10 show the results of our two validation DSNY sub-sections [1021A (Manhattan) and 3114C (Brooklyn)], respectively, and the "Truck route" column represents the individual collection truck associated within a given collection schedule. The overall prediction accuracy of the aggregate total collection by the individual trucks is 99.8% and 93.9% for the two routes, respectively, though we observe wide variations in accuracy of the prediction for specific trucks. These truck-level differences are likely caused by variances in the set-out point for individual buildings, as described above, that lead to the misallocation of the waste from individual buildings to a particular truck. Determining the actual set-out point is a significant operational challenge for DSNY and no data are currently available.

5. Discussion, limitations, and conclusion

The aim of this research is to develop a predictive model for waste generation at the building-level in a dense urban environment, using New York City as a case study. We combine a socio-spatial model of waste generation per capita per week with estimates of the occupant population for each of the more than 750,000 residential buildings in the City. Our best-performing predictive model (GBRT) is able to predict total weekly waste generation for DSNY sub-sections with an out-of-sample R-squared value of 0.87. Subsequent models built predicting refuse, MGP, and paper recycling, respectively, also perform well. We find that the variables with the highest feature importance are weather (temperature, precipitation, wind speed, and snow event), residential building type and density, and demographic variables. Weather-related features capture temporal (i.e. seasonal) variations in the data, in addition to capturing weekly anomalous weather activity. Our building prediction model demonstrates high levels of accuracy following two validation processes. In the two collection truck validation cases, the model resulted in 99.8% and 93.9% prediction accuracy, respectively.

Certain data limitations constrain the predictive power of our model, although iterative improvements are expected as additional validation data are acquired. In order to reflect the spatial

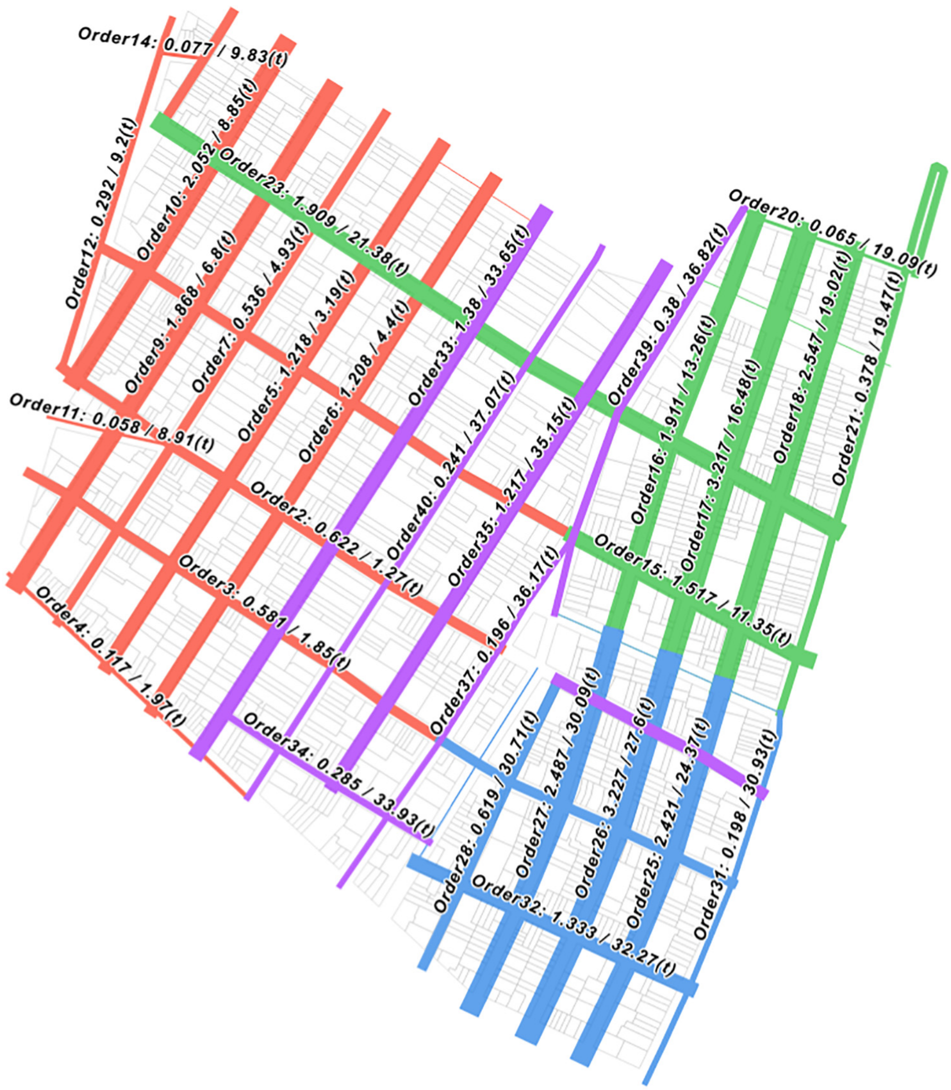

Fig. 12. Truck routes with line width showing relative refuse tonnage (DSNY sub-section 1021A – Monday case).

Table 9
Truck route validation result: Predicted value aggregated for the route vs. actual value (1021A, Monday 2016-10-24, refuse case).

Truck route	Tons collected (predicted)	Tons collected (actual)
MW021M1 (1021A1)	9.83 (t)	13.12 (t)
MW021M2 (1021A2)	11.55 (t)	10.68 (t)
MW021M3 (1021A3)	11.70 (t)	8.70 (t)
MW021M4 (1021A4)	4.48 (t)	5.44 (t)
Total	37.88 (t)	37.94 (t)
Accuracy	99.8%	

Table 10
Truck route validation result: Predicted value aggregated for the route vs. actual value (3114C, Monday 2016-10-24, refuse case).

Truck route	Tons collected (predicted)	Tons collected (actual)
BK114M1 (3114C1)	9.50 (t)	8.69 (t)
BK114M2 (3114C2)	13.14 (t)	10.54 (t)
BK114M3 (3114C3)	7.73 (t)	9.30 (t)
Total	30.37 (t)	28.53 (t)
Accuracy	93.9%	

heterogeneity in waste generation behavior and the propensity of a unit in a building to be occupied, additional features should be considered. Our model could be improved with accurate information on building occupancies at high temporal resolution, particularly accounting for weekly fluctuations in residential population. In addition, specific information on the waste set-out (pick-up/drop-off) point for each building would be useful to match buildings to their true truck route narratives. We are currently working with DSNY to collect these data across selected routes.

The implications of our model are significant. By estimating the amount of waste, refuse, MGP, and paper generated at the building-level with a high degree of accuracy, DSNY can develop more efficient routing schedules for its collection trucks. With annual costs of waste collection reaching hundreds of millions of dollars, even marginal efficiency gains can have non-trivial fiscal impacts. For instance, the number of trips with a partially full truck could be minimized, thus reducing the total vehicle- and person-hours for collection. Similarly, a more efficient allocation of collection trucks could reduce negative air quality impacts and congestion caused by vehicle use and idling. These same models can be applied to other cities around the world, incurring similar savings for other municipalities.

Another implication of our model is our ability to detect areas across the city that have higher or lower propensities to recycle. With this information at a granular level, for individual buildings, DSNY can create targeted outreach programs to promote recycling activity and investigate differences in recycling rates across properties and neighborhoods. These lessons could then be generalized to other cities around the world. Furthermore, a better understanding of the waste generation habits of New York City households can greatly assist city decision-makers in achieving the goal of sending zero waste to landfills by 2030. Data-driven modeling approaches similar to the one created here can help to design and implement more effective, targeted strategies in the future.

Finally, our findings could be used to develop incentive structures to "nudge" households to shift waste behavior. Unit pricing programs, such as "pay-as-you-throw", often raise equity concerns regarding the disproportionate impact of fixed-price unit costs for waste removal for lower-income households, and households that are relatively larger than average. By accurately predicting individual building waste generation rates and accounting for building population and demographic variations, a more equitable pricing structure can be developed. Similarly, our building-level estimates can be used to inform peer comparisons and similar social influence mechanisms to encourage recycling and composting activity. Other incentives, such as demand pricing in high-waste time periods, could be adapted from other utilities, such as the energy sector.

Using machine learning to inform urban operations and planning is a growing research field. Our work adds to the literature both on data-driven city management and urban solid waste management by providing one of the first predictive models of building-level waste and recycling generation. While there are several implications of our model for waste management practice, the methods described here can also have wide application to other urban systems and environmental management more broadly.

Acknowledgements

Our sincere thanks to the New York City Department of Sanitation for providing data, insight, and valuable feedback, as well as to three anonymous reviewers for their constructive comments. We would also like to thank CUSP master's students Geoff Perrin and Xinshi Zheng for their research assistance with preliminary data cleaning. This work has been supported, in part, by the John D. and Catherine T. MacArthur Foundation (Grant number: 16-0875). Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the view of any supporting institution. The authors are responsible for any errors.

References

Adeyemi, A. S., Olorunfemi, J. F., & Adewoye, T. O. (2001). Waste scavenging in third world cities: A case study in Ilorin, Nigeria. *The Environmentalist, 21*, 93–96.

Antanasijevic, D., Pocajt, V., Popovic, I., Redzic, N., & Ristic, M. (2013). The forecasting of municipal waste generation using artificial neural networks and sustainability indicators. *Sustainability Science, 8*, 37–46.

Bhargava, H. K., & Tettelbach, C. (1997). A web-based decision support system for waste disposal and recycling. *Computers, Environment and Urban Systems, 21*, 47–65.

Denafas, G., Ruzgas, T., Martuzevičius, D., Shmarin, S., Hoffmann, M., Mykhaylenko, V., & Ogorodnik, S. (2014). Seasonal variation of municipal solid waste generation and composition in four east European cities. *Resources, Conservation and Recycling, 89*, 22–30.

Dyson, B., & Chang, N.-B. (2005). Forecasting municipal solid waste generation in a fast-growing urban region with system dynamics modeling. *Waste Management, 25*, 669–679.

Esin, T., & Cosgun, N. (2007). A study conducted to reduce construction waste generation in Turkey. *Building and Environment, 42*, 1667–1674.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics,* 1189–1232.

Guerrero, L. A., Maas, G., & Hogland, W. (2013). Solid waste management challenges for cities in developing countries. *Waste Management, 33*, 220–232.

Johnson, N. E., Ianiuk, O., Cazap, D., Liu, L., Starobin, D., Dobler, G., & Ghandehari, M. (2017). Patterns of waste generation: A gradient boosting model for short-term waste prediction in new york city. *Waste Management, 62*, 3–11.

Karadimas, N. V., & Loumos, V. G. (2008). GIS-based modeling for the estimation of municipal solid waste generation and collection. *Waste Management and Research, 26*, 337–346.

Keser, S., Duzgun, S., & Aksoy, A. (2012). Application of spatial and non-spatial data analysis in determination of the factors that impact municipal solid waste generation rates in turkey. *Waste Management,* 359–371.

Kontokosta, C. E. (2015). A market-specific methodology for a commercial building energy performance index. *The Journal of Real Estate Finance and Economics, 51*, 288–316.

Kontokosta, C. E., & Tull, C. (2017). A data-driven predictive model of city-scale energy use in buildings. *Applied Energy, 197*, 303–317.

Langford, M. (2013). An evaluation of small area population estimation techniques using open access ancillary data. *Geographical Analysis, 45*, 324–344.

Langford, M., Maguire, D., & Unwin, D. (1991). *The areal interpolation problem: Estimating population using remote sensing in a GIS framework*. London: Longman.

Leao, S., Bishop, I., & Evans, D. (2004). Spatial–temporal model for demand and allocation of waste landfills in growing urban regions. *Computers, Environment and Urban Systems, 28*, 353–385.

Lwin, K., & Murayama, Y. (2009). A GIS approach to estimation of building population for micro-spatial analysis. *Transactions in GIS, 13*, 401–414.

MacDonald, M. L. (1996). A multi-attribute spatial decision support system for solid waste planning. *Computers, Environment and Urban Systems, 20*, 1–17.

New York City Department of Sanitation.. (2017). *About DSNY*.

New York City Department of Sanitation.. (2017). *NYC Sanitation 2016 Annual Report*.

New York City Mayor's Office of Environmental Coordination.. (2014). *The City Environmental Quality Review technical manual*.

Ojeda-Benítez, S., Armijo-de Vega, C., & Marquez-Montenegro, M. Y. (2008). Household solid waste characterization by family socioeconomic profile as unit of analysis. *Resources, Conservation and Recycling, 52*, 992–999.

Oribe-Garcia, I., Kamara-Esteban, O., Martin, C., Macarulla-Arenaza, A. M., & Alonso-Vicario, A. (2015b). Spatial characteristics of municipal solid waste generation and its influential spatial factors on a city scale: A case study of Xiamen, China. *Waste Management, 39*, 26–34.

Oribe-Garcia, I., Kamara-Esteban, O., Martin, C., Macarulla-Arenaza, A. M., & Alonso-Vicario, A. (2015a). Identification of influencing municipal characteristics regarding household waste generation and their forecasting ability in Biscay. *Waste Management, 39*, 26–34.

Pappu, A., Saxena, M., & Asolekar, S. (2007). Solid wastes generation in India and their recycling potential in building materials. *Building and Environment, 42*, 2311–2320.

Purcell, M., & Magette, W. (2009). Prediction of household and commercial BMW generation according to socio-economic and other factors for the Dublin region. *Waste Management, 29*, 1237–1250.

Reina, V. J., & Kontokosta, C. E. (2017). Low hanging fruit? Regulations and energy efficiency in subsidized multifamily housing. *Energy Policy, 106*, 505–513.

Rimaityte, I., Ruzgas, T., Denafas, G., Racys, V., & Martuzevicius, D. (2012). Application and evaluation of forecasting methods for municipal solid waste generation in an eastern-European city. *Waste Management and Research, 30*, 89–98.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. *Nonlinear estimation and classification* (pp. 149–171). Springer.

Tam, V., & Tam, C. (2006). Evaluations of existing waste recycling methods: A Hong Kong study. *Building and Environment, 41*, 1649–1660.

Thanh, N. P., Matsui, Y., & Fujiwara, T. (2010). Household solid waste generation and characteristic in a Mekong Delta City, Vietnam. *Journal of Environmental Management, 91*, 2307–2321.

The Council of the City of New York.. (2017). *Fiscal 2017 Preliminary Mayor's Management Report for the Department of Sanitation*.

Wang, F., Richardson, A., & Roddick, F. (1996). SWIM–A computer model for solid waste integrated management. *Computers, Environment and Urban Systems, 20*, 233–246.

World Health Organization Centre for Health Development. (2010). *Hidden cities: Unmasking and overcoming health inequities in urban settings.* World Health Organization.

Ye, J., Chow, J.-H., Che, J., & Zheng, Z. (2009). Stochastic gradient boosted distributed decision trees. *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 2061–2064). .

Zade, M. J. G., & Noori, R. (2008). Prediction of municipal solid waste generation by use of artificial neural network: A case study of Mashhad. *International Journal of Environmental Research, 12*, 13–22.

Zhang, G., Lin, T., Chen, S., Xiao, L., Wang, J., & Guol, Y. (2015). Spatial characteristics of municipal solid waste generation and its influential spatial factors on a city scale: A case study of Xiamen, China. *Journal of Material Cycles and Waste Management, 17*, 399–409.