Waste Management 62 (2017) 3–11

Contents lists available at ScienceDirect

Waste Management

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

Patterns of waste generation: A gradient boosting model for short-term waste prediction in New York City

Nicholas E. Johnson [a,b,*], Olga Ianiuk [a], Daniel Cazap [a], Linglan Liu [a], Daniel Starobin [c], Gregory Dobler [a], Masoud Ghandehari [a]

[a] Center for Urban Science and Progress, New York University, New York, NY 11201, United States
[b] Warwick Institute for the Science of Cities, University of Warwick, Coventry CV47AL, United Kingdom
[c] New York City Department of Sanitation, New York, NY 10007, United States

ARTICLE INFO

Article history:
Received 10 October 2016
Revised 4 January 2017
Accepted 25 January 2017
Available online 16 February 2017

Keywords:
Waste
Gradient boosting
Prediction
New York City

ABSTRACT

Historical municipal solid waste (MSW) collection data supplied by the New York City Department of Sanitation (DSNY) was used in conjunction with other datasets related to New York City to forecast municipal solid waste generation across the city. Spatiotemporal tonnage data from the DSNY was combined with external data sets, including the Longitudinal Employer Household Dynamics data, the American Community Survey, the New York City Department of Finance's Primary Land Use and Tax Lot Output data, and historical weather data to build a Gradient Boosting Regression Model. The model was trained on historical data from 2005 to 2011 and validation was performed both temporally and spatially. With this model, we are able to accurately ($R2 > 0.88$) forecast weekly MSW generation tonnages for each of the 232 geographic sections in NYC across three waste streams of refuse, paper and metal/glass/plastic. Importantly, the model identifies regularity of urban waste generation and is also able to capture very short timescale fluctuations associated to holidays, special events, seasonal variations, and weather related events. This research shows New York City's waste generation trends and the importance of comprehensive data collection (especially weather patterns) in order to accurately predict waste generation.

1. Introduction

With the rapid development of urban environments around the world, municipal waste generation is fast becoming one of the most pressing issues facing cities globally. Currently, at 3.3 million tons per day, the global production of waste is already becoming unmanageable, and this rate is expected to grow to 11 million tons per day by 2100 (Hoornweg et al., 2013). Given these trends, effective urban waste management systems are essential, and in order to provide these services in an environmentally sound and financially sustainable way, there is an urgent need for basic understanding of the amount and composition of the materials produced (Beigl et al., 2008; Rimaityte et al., 2012). Furthermore, forecasting waste generation becomes a critical aspect of urban waste management that provides city agencies the ability to optimize collection and disposal operations in the short term, as well as develop long-term strategies for disposal planning, policy development, and implementation of waste reduction programs (Chang and Lin, 1997). The goal of this research is to use historical municipal solid waste (MSW) data supplied by the New York City Department of Sanitation (DSNY) to forecast waste generation across the city.

2. Modeling of waste data

A variety of modeling methodologies have been used to forecast waste generation including the use of group comparison, correlation analysis, multiple regression analysis, input-output analysis, time-series analysis, and system dynamics modeling (Beigl et al., 2008). These models often focus on identifying the underlying relationship between variables that drive waste generation. For example, at the municipal level, Oribe-Garcia et al. (2015) identified urban morphology, tourism activity, level of education, and income as the most influencing factors leading to MSW generation while Daskalopoulos et al. (1998) used single regression analysis to link gross domestic product and related total consumer expenditure as strong correlating factors in waste generation at the country level. Navarro-Esbri et al. (2002) and Rimaityte et al. (2012) used traditional time-series approaches such as Autoregressive and

* Corresponding author at: Center for Urban Science and Progress, New York University, New York, NY 11201, United States.
E-mail address: Nicholas.johnson@nyu.edu (N.E. Johnson).

http://dx.doi.org/10.1016/j.wasman.2017.01.037
0956-053X/

Integrated Moving Average (ARIMA) and seasonal Autoregressive and Integrated Moving Average (sARIMA) to predict generation. Xu et al. (2013) disregarded demographic and socioeconomic factors and forecasted waste generation using a hybrid sARIMA model and grey system theory, a methodology to reveal the dynamic relationships in a system using differential equations that is derived from control theory in which the term grey describes the understanding of information in the system (a system is defined as "grey" if the information about the system is only partially known). Other approaches to model waste generation include Zade and Noori (2007) who used artificial neural networks to predict weekly waste generation in Tehran and Abbasi et al. (2012) who used partial least squares for feature selection and support vector machines to predict for the same area.

Our work builds upon these data-centric forecasting approaches in several ways. First, the close collaboration with the DSNY provided detailed information about citywide operations and the specific challenges faced by the agency. The agency's expert knowledge of the city's waste system was important for proper data organization and cleaning processes, including specific information about how source data was generated, as well as provided insight on and confirmation of preliminary analyses. Second, the breadth and depth of the historical data provided by the DSNY is unprecedented in urban waste forecasting studies. Not only is this dataset highly granular both temporally and spatially, it also spans a full ten-years that allows for robust statistical results and thorough model cross-validation. Finally, this research uniquely uses a Gradient Boosting Regression model for forecasting in both time and space for New York City.

3. Waste in NYC

Currently all of NYC's refuse is exported out of the city through a network of contract vendors. These vendors use a combination of long-haul trailer trucks (48%), trains (42%), and direct haul (10%). At present, 80% of the city's solid waste is disposed of in landfills located in New York, Pennsylvania, Ohio, South Carolina, Virginia, and Kentucky, and 20% is disposed of in waste-to-energy facilities in New York, New Jersey, Pennsylvania, and Connecticut. The operational budget for the DSNY in 2012 was $1.6 billion dollars (Kellerman and Gibbs, 2014).

In 2007, NYC Mayor Michael Bloomberg launched PlaNYC that established the goal of diverting 75% of the city's solid waste from landfills by 2030 (Bloomberg, 2006). To achieve this goal, the DSNY established a pilot organic collection program to capture food scraps, yard clippings and soiled paper that serviced 100,000 homes and 40% of NYC schools as well as enhanced drop-off programs for diverting other waste including textiles, e-waste and household hazardous materials. The recycling program was expanded in 2013 as a result of the construction of a new state-of-the art facility which allowed the collection of all rigid plastics for recycling.

In 2015, Mayor de Blasio announced the OneNYC plan that, among other initiatives, sets a citywide goal of 90% reduction of waste disposed in landfills by 2030 (de Blasio, 2015). To achieve this goal, the city aims to expand its organics collection program to the entire city, create a single-stream recycling program to enhance curb-side collection, expand the recycling program to include New York City Housing Authority buildings, reduce the use of non-compostable wastes (plastic bags) and initiate zero waste programs in NYC schools.

3.1. DSNY data

To manage the waste generated by NYC's ~8.4 million inhabitants, DSNY employs over 7000 sanitation workers, servicing 59 community districts. In total, 12,000 tons of residential refuse and recycling is produced in the city each day. The agency's purview includes collection of waste from city residents, public agencies and not-for-profit organizations as well as street cleaning and snow clearing mandates. DSNY collects ~25% of the total waste produced in NYC. The remaining 75% is handled by private haulers and includes commercial/business waste, construction and demolition waste, and industrial waste.

DSNY provides bi-weekly or tri-weekly MSW collection services throughout the city as well as recycling collection once per week. NYC residents are required to separate recyclables into two separate bins, one comprised of metal, glass and plastic (MGP) and the other with mixed cardboard and paper. The city's recycling program began in 1989 though it was suspended for two years from 2002 to 2004.

3.2. Data collection

The waste collection data provided by DSNY spans more than a decade, from July 2003 to January 2015. Each record in this dataset contains the collection information for a single truck. Specifically, each record holds a unique truck ID, the collected tonnage inferred by weighing the truck, the time the truck was weighed, the type of material collected (refuse, metal/glass/plastic, paper), and the geospatial area from which the waste was collected (DSNY uses 232 geographies across NYC's five boroughs that are referred to as sections).

Fig. 1 shows the total weekly collection tonnage integrated across all sections of the city for both refuse and recycling. There are clear temporal patterns at multiple timescales. For example, strong seasonality is apparent with higher waste generation rates during the summer and lower generation rates during the winter. This observation is consistent with previous research (Korhonen and Kaila, 2015; Denafas et al., 2014). A decreasing trend in refuse generation can also be identified: in 2005, the average weekly refuse generation was approximately 60,000 tons, which has slowly declined to 48,000 tons per week in 2014 despite an increasing urban population. This overall decline in waste generation is in part due to the effects of the recession in 2008 as well as reduced product size and packaging (Garcia, 2014). New York City's recycling program initiated in 2004 and has maintained an average recycling rate of 17%.

In January of 2011, a winter storm produced 20 in. of snow resulting in a significant aberration in waste collection, visible by the sharp increase and decrease in tonnage. The snowfall paralyzed the city's means of transportation and left vehicles abandoned roadside reducing DSNY's ability to clear snow. Subsequently, waste collection was delayed across the city. This interruption is visible in the sharp decline in tonnage during the event and an increase in tonnage following the event when DSNY began collecting refuse and recycling that accumulated during snow clearing operations.

Recycling rates vary across DSNY collection sections. Fig. 2 shows the distribution of daily per capita waste generation that highlights a narrow distribution for the refuse stream ranging from 1 lb to 2.5 lb, while the paper and MGP recycling streams show a much larger distribution. Clarke and Maantay (2006) suggest that the variance in recycling rates in New York City are strongly correlated to four variables including percentage of persons below the poverty level, percentage of households headed by a single female with children, percentage of adults without a high school diploma and the percent of minority population.

4. Methods

Figs. 1 and 2 demonstrate that there is tremendous complexity in the data gathered by DSNY, though there are also clear patterns

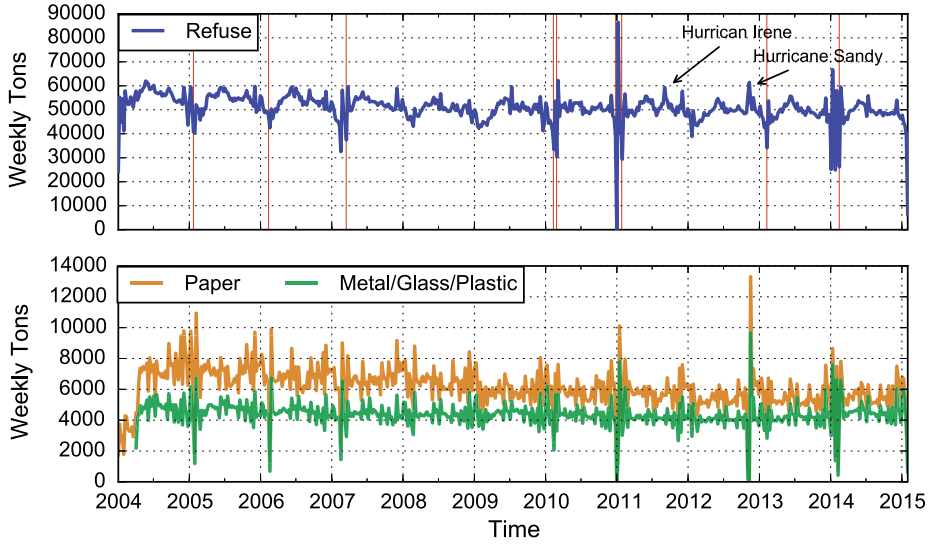

Fig. 1. New York City's weekly refuse and recycling collection tonnages from 2004 to 2015. The vertical red lines indicate significant winter weather events with above average snowfall.

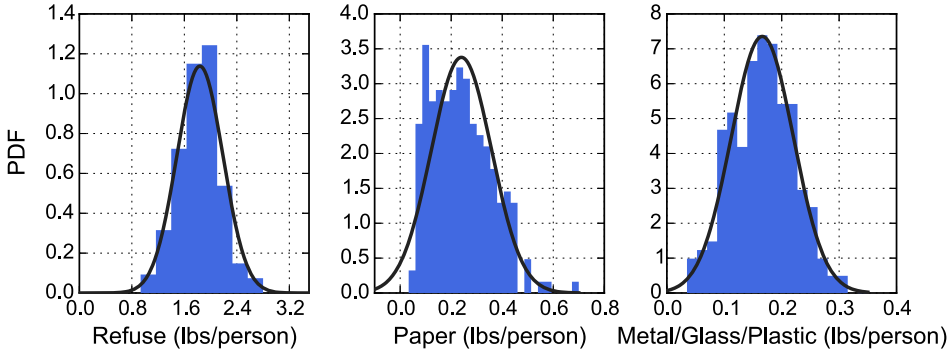

Fig. 2. Probability density function of daily per capita waste and recycling generation for DSNY sections. The black lines indicate a normal distribution.

that can be extracted for the purpose of making near-term (~1–2 weeks in advance) spatiotemporal predictions of tonnages. As noted above, such predictions would be useful for both day-to-day operations and long term planning. We now present a machine learning (ML) model designed to make such predictions using information from the data stream itself as well as external temporal data sets.

4.1. Model

A Gradient Boosted Regression Tree (GBRT) model was chosen to predict short-term waste generation. GBRT is based on decision tree regression models that are often used in ML because of their interpretability, ability to handle non-linear and complex relationships between data and have demonstrated higher prediction accuracy compared to traditional time-series models such as ARIMA (Oliveira and Torgo, 2014; Kane et al., 2014). Decision trees, described by Breiman et al. (1984) and Hastie et al. (2001), use binary splitting to divide a feature space into regions and fit a different (linear) model to each region. This process is performed recursively and at each stage, the split point is determined by the greatest reduction in residual sum of squares (RSS). The process results in a single tree-like structure that "best" describes the underlying relationship between variables in a dataset.

GBRT extends traditional decision tree modeling by incorporating a statistical technique called boosting. Instead of building one "best fit" tree model, boosting improves prediction accuracy by building many "weak" models that are then aggregated to form a single consensus model (Schapire, 2003). Weak models are defined as performing only slightly better than random guessing. Using GBRT, decision trees are created sequentially using the residuals of the first tree as the input of the new tree. In this manner, the model learns the relationship between features based on the errors of the previous tree. This sequential model building process is a form of functional gradient descent that optimizes prediction by adding a new tree at each stage that best reduces the loss function (Elith et al., 2008).

Several parameters are tuned to optimize model performance. The number of trees and their depths control the final tree's structure and complexity. Tree depth is particularly significant because it determines the degree of interaction between features. Since trees are grown sequentially, each new tree takes into account the previous trees and therefore shallow trees with a depth of 4–6 are often preferred (Friedman, 2001). The model's learning rate is another important parameter that determines how much each tree contributes to the overall model. A low learning rate will increase the number of trees used, which is ideal for better performance, but requires an increase in computation time. The final model is a linear combination of all decision trees whose contribution to the overall model is weighted by the shrinkage parameter (learning rate). GBRT was implemented using Python's Scikit-learn package version 1.17 (Pedregosa et al., 2012).

Two GBRT models were built. A spatiotemporal model was built whose features include spatial and temporal characteristics, while a second, time-integrated spatial model was built in order to validate the spatial relationships between features. The spatiotemporal model was trained using weekly data from 2005 to 2011 and weekly predictions were made for each of the 232 DSNY sections for the year 2012. Validation was performed following each weekly prediction and R2 scores were averaged over the 52 prediction weeks. After each prediction iteration, the actual tonnage from the predicted week was then incorporated into the training set for the next iteration. The number of trees built for each model was chosen to be 1000. The model's learning rate was set to 0.1 and the maximum number of features in each tree was 6.

The spatial model was built ignoring the temporal aspect of the data. This was accomplished by training the model on data from 192 randomly selected sections (~83%) and predicting for the remaining 40 sections. 10-fold cross-validation was performed and results were averaged. Model parameters were identical to the temporal model.

Metrics for model performance were calculated using Root Mean Squared Error (RMSE), and the coefficient of determination (R2). These metrics were computed for each prediction iteration where each iteration predicted a weekly tonnage for each of the 232 DSNY sections. The weekly metrics were then averaged over the 52-week study period.

The working models include 28 features and predictions were made across three different waste streams including refuse, paper recycling and MGP recycling. Each model was also run three separate times with different feature groups. The first iteration used all features while the second iteration used only internal features and third iteration used only external features (see Section 4.2). For the prediction week, weather averages, holidays and disaster notifications are assumed from weekly forecasts.

4.2. Feature development

Feature selection resulted in the development of two groups of features; internal and external features. Internal features are generated from the DSNY waste dataset while external features are generated from alternate data sources. Features were selected either because of their strong relationship to waste generation as indicated through previous research or because of their overall temporal and/or spatial granularity and significance in regards to the study period and area.

Fig. 3 shows the autocorrelation of tonnage as a function of time for lags between 0 and 80 weeks. The high correlation at 52 weeks and 4 weeks indicate strong seasonal and annual waste generation patterns and suggest several internal features useful for our GBRT model: the previous year's tonnage, the tonnage four weeks prior, and the previous week's tonnage.

External features are generated from outside of the supplied DSNY dataset. Demographic and socioeconomic data were obtained from the 2010 American Community Survey and the Longitudinal Employer-Household Dynamics (LEHD) data that include three age brackets, three income brackets, five race groups, and four categories of educational attainment. The LEHD dataset includes annual socioeconomic and demographic data at the census block level from 2009 to 2012. Other external features were included for their significant temporal and/or spatial granularity. These data include historical weather and data derived from the New York City's Department of Finance Primary Land Use and Tax Lot Output (PLUTO). Features included from PLUTO were the average lot market value, the total number of residential units, population density (population divided by the total residential area), and the percent of residential lots.

Information about National American holidays was also included due to the noticeable peaks following July 4th, December 25th, etc. Table 1 provides a complete list of features.

Many considerations were taken in aligning these datasets given the diverse temporal and spatial nature of the data. Temporally, in order to appropriately align the multiple datasets, only data from years 2005 through 2012 were used. The supplied DSNY data was also aggregated to the week level generating a total weekly tonnage for individual geographies across the city. The DSNY does not provide daily collection for each household and at most, collection occurs tri-weekly. Therefore, the total weekly collection tonnage was chosen as a common temporal scale in order to compare different sections across the city. This aggregation also mitigates the numerous fluctuations and variations that occur within individual sections and makes waste generation patterns apparent at the section level. Weather data included average weekly temperature, average weekly humidity, the weekly precipitation total and the number of days with weather events.

Spatial data aggregation was also necessary in some instances. Spatial areas for ACS, LEHD and PLUTO datasets exist at a more granular spatial resolution than the DSNY waste data. In these cases, data were joined spatially to match the DSNY spatial dimensions and the sum or mean value was computed. It was observed that the granular spatial dimensions of ACS, LEHD and PLUTO were contained within the DSNY sections and there were no partial contained geometries. Fig. 4 is a map of Lower Manhattan and showcases the spatial dimensions of a typical DSNY section and the census and PLUTO sub-units.

5. Results and discussion

5.1. Model performance

Table 2 shows the average R2 and RMSE results for both the spatiotemporal model and the spatial model for each waste stream. Overall spatiotemporal model performance shows good prediction accuracy for all waste streams, though performance varies per stream and per feature group (i.e. internal features only, external features only, all features). The results show that prediction for the refuse stream consistently performs better than the other two waste streams with approximately 15% increase in accuracy over paper predictions and a 20% increase over MGP predictions. When using all features, the average refuse prediction accuracy is 88% while the average prediction accuracy for paper and MGP streams are 74% and 68% respectively. Spatiotemporal model performance using only internal features performs better than using only external features for each waste stream. Similar to the spatiotemporal model, the results for the spatial model also show good prediction accuracy for refuse, though decrease for the recycling streams.

Fig. 5 shows refuse prediction results for the spatiotemporal model using multiple feature groups. The weekly tonnage results are aggregated from individual predictions at the DSNY section level to demonstrate waste generation predictions citywide. Fig. 6 shows refuse prediction results from the spatiotemporal model for DSNY section MN032 located in lower Manhattan.

5.2. Feature importance

Fig. 7 highlights the top 20 features ranked by their relative feature importance for refuse prediction using the spatiotemporal and spatial models. Scores were averaged over all model iterations. Relative feature importance is an indicator of a feature's contribution in predicting a target response and is determined by how often a feature is used in the split points of a tree. The more often a feature

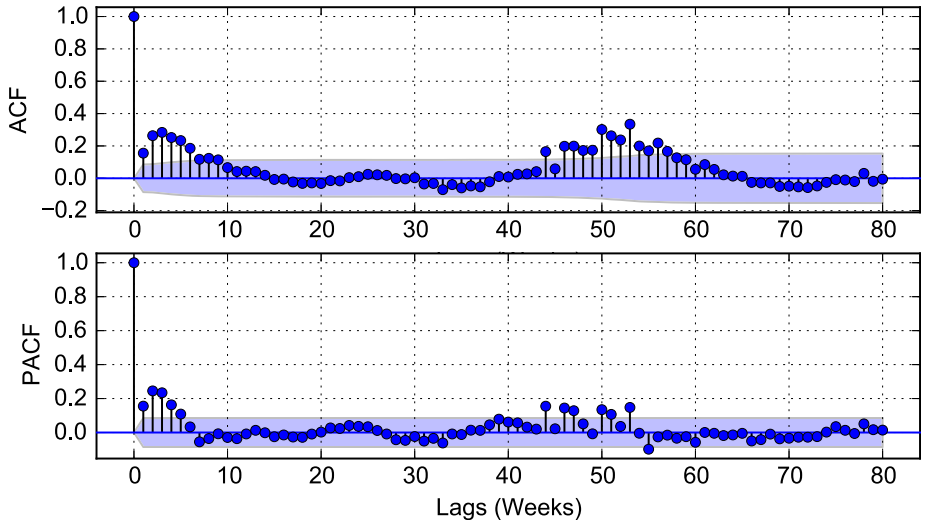

Fig. 3. Autocorrelation and partial autocorrelation with 95% confidence intervals shown in light blue where the standard deviation is computed using Bartlett's formula. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Complete list of features used for prediction.

Data source	Timescale	Feature description
DSNY	Weekly (2005–2012)	Week minus 52 weeks Week minus 4 weeks Week minus 1 week
PLUTO	Static	Residential landuse percentage[a] Number of residential units Population density[b] Average total lot value
Weather underground	Weekly	Average weekly temperature Total weekly precipitation Average weekly humidity Weather events[c]
Data.gov	Static	National Holidays
LEHD	Annual (2009–2012)	Age 29 or younger Age 30–54 Age 55 or older Earning $1250/month or less Earning $1251/month to $3333/month Earning greater than $3333/month Race: White Race: Black or African American Race: Asian Race Native Hawaiian or other Pacific Islander Education: Less than high school Education: High school equivalent, no college Education: Some college or Associate Degree Education: Bachelor degree or advanced degree

[a] Landuse categories 01, 02, 03, 04.
[b] Section population divided the number of section residential units.
[c] Binary sum of hourly weather observations (rain, snow, fog, etc.).

is used, the greater the feature's importance. For ensemble tree models, values are averaged across the trees for each feature and the sum of all feature importances is equal to 1 (Friedman, 2001).

The relative feature importance for both the spatiotemporal model and the spatial model indicates the same top ten features are the most significant features used in the models and comprise approximately 90% of the relative feature importance. Similarly, the same top five features collectively account for approximately 75% of the relative feature importance. All three internal features rank amongst the top five features, though the most significant feature used in the model is temperature. All weather features are included in the top ten relative feature importance. It should also be noted that demographic and socio-economic features are largely uninformative for short-term prediction. As seen in Fig. 7, only income (<$1250/month) and age (>55) appear in the top ten list of important features contributing approximately 3% and 1% to the model's predictive power respectively.

6. Discussion

Overall, the model demonstrates an ability to predict waste generation with highly granular spatial and temporal accuracy across multiple waste streams. The best model performance however, occurs when predicting refuse compared to predicting either recycling stream. We conclude this results from the selected features used in the model, granular spatial and temporal data for model training, and most importantly, the regular nature of refuse generated in New York City. This regularity is observed in both space and time and persists through disruptions in collection caused by extreme events. In the case of extreme events, every under-collection caused by a disruptive weather event must result in a corresponding "make-up" collection following shortly thereafter in order to avoid a buildup of uncollected waste. For example, this can be seen in Fig. 1 where sharp declines in collection tonnages that can be explained by exceptional events (e.g., winter storms) that disrupted DSNY collection efforts are immediately followed by an increase in collection following the event. This demonstrates that while the DSNY collection efforts were interrupted, public waste generation continued. This furthermore supports the argument that the data aggregated to the week more accurately represents the regularity and steady flow of waste generated by society.

Figs. 5 and 6 show highly accurate prediction results throughout the year with the exception of two extreme weather events including an above average snowfall in late January and Hurricane Sandy in late November. The nature of these events make them extremely difficult to predict and overall prediction accuracy decreases during these events. However, by including weather information, the model does have the ability to rebound from these events.

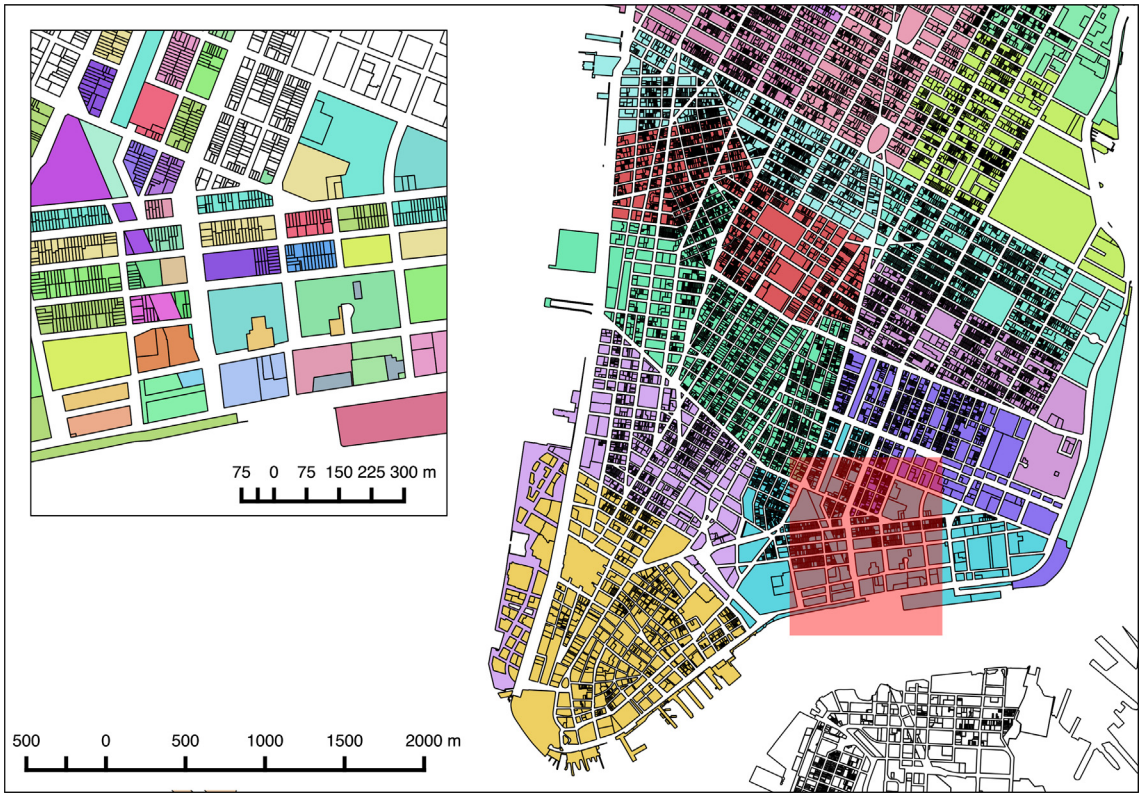

Fig. 4. Department of Sanitation waste collection map. Large multi-colored areas indicate DSNY sections while smaller multi-colored areas in the inset map depict census blocks.

Table 2
Prediction results.

Stream	Group	R2		RMSE	
		Spatial	Spatiotemporal	Spatial	Spatiotemporal
Refuse	All features	0.906	0.889	22.059	21.632
	External	0.604	0.837	45.326	28.560
	Internal	0.871	0.875	25.812	22.751
Paper	All features	0.791	0.744	6.950	6.337
	External	0.628	0.508	9.295	8.670
	Internal	0.738	0.731	7.779	6.435
MGP	All features	0.694	0.685	4.222	4.320
	External	0.428	0.578	5.772	5.157
	Internal	0.606	0.658	4.790	4.525

Another important observation relating to extreme weather is the impact of historical weather on current predictions. Neither model directly includes historical weather data beyond historical data used for training purposes. Each prediction is made considering only the weather for the prediction week. However, each prediction does include the previous year's tonnage, which in some cases, is affected by previous extreme weather events. For example, as previously mentioned an above average snowfall took place in late January of 2012. After investigation, it was observed that a similar extreme winter event took place the previous week exactly one year earlier. From the model's perspective, the previous storm impacted the previous year's tonnage, which is a feature included for prediction and can therefore significantly alter prediction results. These fluctuations can be observed in Fig. 5 during the last two weeks of January and the beginning of February.

Feature importance results provide insight into the model's decision-making process but should not be interpreted as causality or the strength of the relationship between the dependent and independent variables. For both spatial and spatiotemporal models, temperature is ranked as the most significant feature. One interpretation is that temperature contains the high frequency time-scale information that captures weekly fluctuations as well as provides seasonal information capturing the low frequency patterns in refuse generation as previously observed in the autocorrelation. A comparison of models using internal features versus external features clearly shows that models using internal features perform better than models with external features only. One can conclude that while weather is an important element for prediction providing high and low frequency information, the inclusion of waste data yields a better model for prediction. It should also be noted that demographic and socioeconomic features are largely uninformative for short-term prediction. As seen in Fig. 7, only income (<$1250/month) and age (>55) appear in the top ten list of important features contributing approximately 3% and 1% to the model's predictive power respectively.

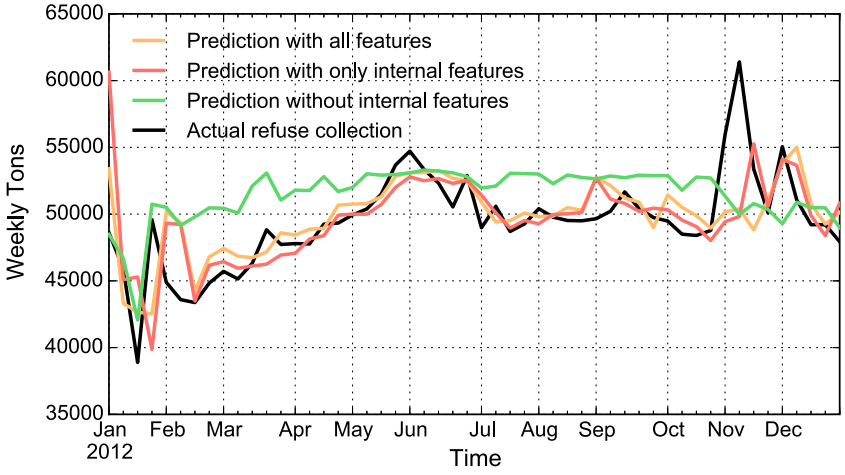

Fig. 5. Spatiotemporal refuse prediction integrated for all sections using multiple feature groups.

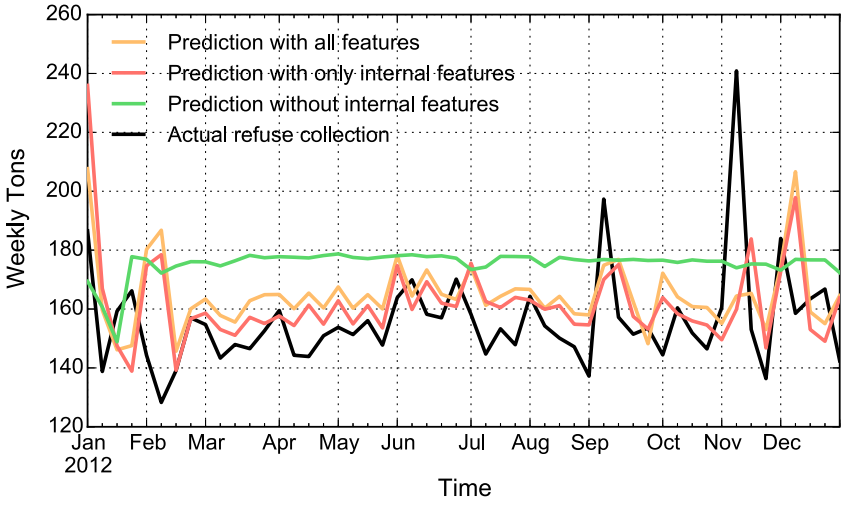

Fig. 6. Example of spatiotemporal refuse prediction for an individual DSNY section (MN032) located in lower Manhattan.

Another important observation is the higher R2 scores for spatial models compared to spatiotemporal models. The high R2 scores for the spatial model validates the spatial features and indicates that space is an important element for prediction. If the spatial features were insignificant, one would expect poor model performance. However, while the spatial model does produce a higher R2 score, it is important to consider the RMSE of each model. Indeed, the spatiotemporal model consistently predicts with a lower RMSE, highlighting sensitivity between the two models that makes direct comparison difficult. One explanation for this is the methodology used for each model in which the spatial model predicts waste for 40 sections over 416 weeks for each iteration, while the spatiotemporal model predicts 232 sections for one week each iteration. Because results are averaged, the spatiotemporal model is therefore more sensitive to large errors often resulting from extreme weather events.

As previously noted, a key insight to this work is the identification of regular waste generation patterns. These patterns, though identified specifically for New York City, are likely to exist with varying degrees across various spatiotemporal scales around the world, which suggests this work could be applied to other cities and regions depending on the availability and granularity of historical waste data. Accurate long-term predictions could also be achieved through the use or aggregation of data to various temporal scales. Indeed, long-term (annual) predictions using granular spatial and temporal data may provide similar results and provide greater insight on the relationship between slow-changing socioeconomic and demographic features that largely provide little short-term predictive ability.

Lastly, we note that the high granularity in space and time of the DSNY data provides the most significant predictive power needed for forecasting waste generation (and indeed, our results show that historical waste data alone is sufficient in all but the most extreme cases). With only one exception, the spatiotemporal model predicting refuse using external features only, the absence of waste data consistently results in poor model performance. The effort and ability of the DSNY to collect data about waste collection serves as a strong example for other cities aiming to improve the understanding and operations through data-driven analysis. The collection and integration of data is fundamental to enable data-driven decisions for enhanced operations, improved planning and ultimately for better policy decision-making.

7. Conclusions

We have developed a short-term predictive model for waste generation in New York City. Multiple datasets were integrated

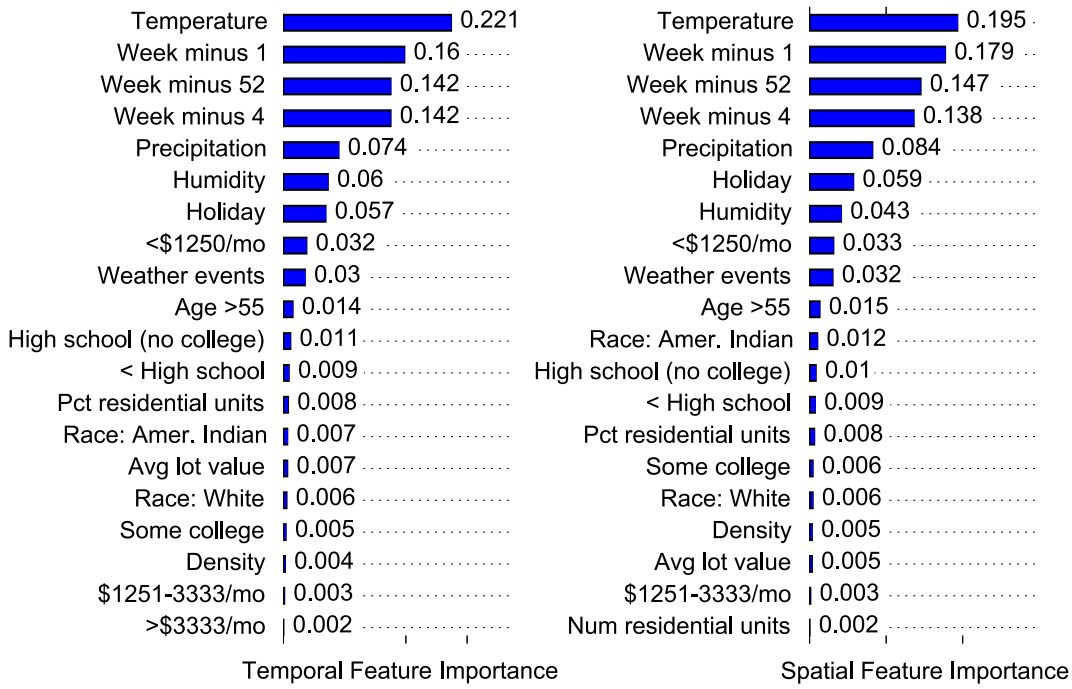

Fig. 7. Importance ranking for features used in spatial and spatiotemporal model predicting refuse.

and combined with features from historical waste collection data by the New York City Department of Sanitation. A gradient boosting regression model was built to predict weekly waste generation across three streams with an average accuracy of 88%. We have shown that the success of the model is largely due to the fine temporal and spatial granularity of the DSNY data.

The model results demonstrate that refuse generation is predictable across the city and the ability to predict waste generation can largely be explained by the regularity of waste generation by society when aggregated to one week timescales. This regularity allowed us to derive features based purely on the waste tonnage data alone that provide excellent prediction accuracy. Furthermore, this temporal behavior is similar across spatial regions although the absolute amount of waste in different regions varies significantly. Including additional, external features (most importantly weather conditions) further improved the robustness of the model.

This research was done in collaboration with the New York City Department of Sanitation in order to improve operational efficiencies for both short-term and long-term planning. The ability to forecast waste generation gives the DSNY a unique ability to optimize waste collection and vehicle allocation as well as the potential to develop targeted long-term strategies for waste reduction and recycling programs. This research also offers a foundation for city agencies to begin experimenting and understanding the effectiveness of waste management programs in various parts of the city.

Acknowledgements

This research is supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) Centre for Doctoral Training in Urban Science and Progress (EP/L016400/1).

References

Abbasi, M., Abduli, M., Omidvar, B., Baghvand, A., 2012. Forecasting municipal solid waste generation by hybrid support vector machine and partial least square model. Int. J. Environ. Res. 7, 27–38.

Beigl, P., Lebersorger, S., Salhofer, S., 2008. Modeling municipal solid waste generation: a review. Waste Manage. 28, 200–214. http://dx.doi.org/10.1016/j.wasman.2006.12.011. <http://www.ncbi.nlm.nih.gov/pubmed/17336051>.

Bloomberg, M., 2006. A Greener Greater New York. The City of New York.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA, USA.

Chang, N.B., Lin, Y.T., 1997. An analysis of recycling impacts on solid waste generation by time series intervention modeling. Resour. Conserv. Recycl. 19, 165–186. http://dx.doi.org/10.1016/S0921-3449(96) 01187-1.

Clarke, M.J., Maantay, J.A., 2006. Optimizing recycling in all of New York City's neighborhoods: using GIS to develop the reap index for improved recycling education, awareness, and participation. Resour. Conserv. Recycl. 46, 128–148.

Daskalopoulos, E., Badr, O., Probert, S., 1998. Municipal solid waste: a prediction methodology for the generation rate and composition in the European Union countries and the United States of America. Conserv. Recycl. 24, 155–166. http://dx.doi.org/10.1016/S0921-3449(98) 00032-9. <http://www.sciencedirect.com/science/article/pii/S0921344998000329>.

Denafas, G., Ruzgas, T., Martuzevicius, D., Shmarin, S., Hoffmann, M., Mykhaylenko, V., Ogorodnik, S., Romanov, M., Neguliaeva, E., Chusov, A., Turkadze, T., Bochoidze, I., Ludwig, C., 2014. Seasonal variation of municipal solid waste generation and composition in four East European cities. Resour. Conserv. Recycl. 89, 22–30. http://dx.doi.org/10.1016/j. resconrec.2014.06.001.

de Blasio, B., 2015. One New York: The Plan for a Strong and Just City. Technical Report, The City of New York. <http://nyc.gov/onenyc>.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A Working Guide to Boosted Regression Trees. doi:http://dx.doi.org/10.1111/j.1365-2656.2008.01390.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232. http://dx.doi.org/10.1017/CBO9781107415324.004. Available from: arXiv: 1011.1669v3.

Garcia, K., 2014. Local Law 77 of 2013 Organics Collection Pilot Program. Technical Report Department of Sanitation.

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag, New York.

Hoornweg, D., Bhada-Tata, P., Kennedy, C., 2013. Waste production must peak this century. Nature 502, 615–617.

Kane, M.J., Price, N., Scotch, M., Rabinowitz, P., 2014. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinform. 15 (1), 276. http://dx.doi.org/10.1186/1471-2105-15-276.

Kellerman, C., Gibbs, K., 2014. 12 Things New Yorkers Should Know About Their Garbage. Technical Report Citizen Budget Commission.

Korhonen, P., Kaila, J., 2015. Waste container weighing data processing to create reliable information of household waste generation. Waste Manage. 39, 15–25. http://dx.doi.org/10.1016/j.wasman.2015.02.021.

Navarro-Esbri, J., Diamadopoulos, E., Ginestar, D., 2002. Time series analysis and forecasting techniques for municipal solid waste management. Resour. Conserv. Recycl. 35, 201–214. http://dx.doi.org/10.1016/S0921-3449(02)00002-2.

Oliveira, M., Torgo, L., 2014. Ensembles for time series forecasting. In: ACML.

Oribe-Garcia, I., Kamara-Esteban, O., Martin, C., Macarulla-Arenaza, A.M., Alonso-Vicario, A., 2015. Identification of influencing municipal characteristics regarding household waste generation and their forecasting ability in Biscay. Waste Manage. 39, 26–34. http://dx.doi.org/10.1016/j.wasman. 2015.02.017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2012. Scikit-learn:

machine learning in python. J. Mach. Learn. Res. 12, 2825–2830. http://dx.doi.org/10.1007/s13398-014-0173-7.2. Available from: arXiv: 1201.0490. <http://dl.acm.org/citation.cfm?id=2078195$\delimiter"026E30F$n>.

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 Manage. Res.: J. Int. Solid Waste. Public Cleans. Assoc. ISWA 30, 89–98. http://dx.doi.org/10.1177/0734242X10396754. <https://www.ncbi.nlm.nih.gov/pubmed/21382880>.

Schapire, R., 2003. The boosting approach to machine learning – an overview. In: Denison, D.D., Hansen, M.H., Holmes, C., Mallick, B., Yu, B. (Eds.), MSRI Workshop on Nonlinear Estimation and Classification 2002. Springer, New York.

Xu, L., Gao, P., Cui, S., Liu, C., 2013. A hybrid procedure for MSW generation forecasting at multiple time scales in Xiamen City, China. Waste Manage. 33, 1324–1331. http://dx.doi.org/10.1016/j.wasman.2013.02.012.

Zade, M.J.G., Noori, R., 2007. Prediction of municipal solid waste generation by use of artificial neural network: a case study of Mashhad. Int. J. Environ. Res. 2 (1) (ISSN: 1735-6865).