INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Abdolhosseini Qomi MJ, Noshadravan A, Sobstyl JM, Toole J, Ferreira J, Pellenq RJ-M, Ulm F-J, Gonzalez MC. 2016 Data analytics for simplifying thermal efficiency planning in cities. *J. R. Soc. Interface* **13**: 20150971.
http://dx.doi.org/10.1098/rsif.2015.0971

Received: 7 November 2015
Accepted: 21 March 2016

Subject Category:
Life Sciences–Engineering interface

Subject Areas:
environmental science, mathematical physics

Keywords:
massive–passive data analytics, strategic gas consumption planning, probabilistic model reduction, response surface methodology

Author for correspondence:
Marta C. Gonzalez
e-mail: martag@mit.edu

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2015.0971 or via http://rsif.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

Data analytics for simplifying thermal efficiency planning in cities

Mohammad Javad Abdolhosseini Qomi[1], Arash Noshadravan[2], Jake M. Sobstyl[3], Jameson Toole[4], Joseph Ferreira[5], Roland J.-M. Pellenq[3,6,7], Franz-Josef Ulm[3,5] and Marta C. Gonzalez[3,4]

[1]Department of Civil and Environmental Engineering, University of California at Irvine, Irvine, CA 92617, USA
[2]Zachary Department of Civil Engineering, Texas A&M University, TX 77843, USA
[3]Department of Civil and Environmental Engineering, [4]Engineering Systems Division, [5]Department of Urban Studies and Planning, and [6]MSE2 MIT-CNRS Joint Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[7]Centre Interdisciplinaire des Nanosciences de Marseille, CNRS and Marseille Université, Campus de Luminy, Marseille, 13288 Cedex 09, France

More than 44% of building energy consumption in the USA is used for space heating and cooling, and this accounts for 20% of national CO_2 emissions. This prompts the need to identify among the 130 million households in the USA those with the greatest energy-saving potential and the associated costs of the path to reach that goal. Whereas current solutions address this problem by analysing each building in detail, we herein reduce the dimensionality of the problem by simplifying the calculations of energy losses in buildings. We present a novel inference method that can be used via a ranking algorithm that allows us to estimate the potential energy saving for heating purposes. To that end, we only need consumption from records of gas bills integrated with a building's footprint. The method entails a statistical screening of the intricate interplay between weather, infrastructural and residents' choice variables to determine building gas consumption and potential savings at a city scale. We derive a general statistical pattern of consumption in an urban settlement, reducing it to a set of the most influential buildings' parameters that operate locally. By way of example, the implications are explored using records of a set of ($N = 6200$) buildings in Cambridge, MA, USA, which indicate that retrofitting only 16% of buildings entails a 40% reduction in gas consumption of the whole building stock. We find that the inferred heat loss rate of buildings exhibits a power-law data distribution akin to Zipf's law, which provides a means to map an optimum path for gas savings per retrofit at a city scale. These findings have implications for improving the thermal efficiency of cities' building stock, as outlined by current policy efforts seeking to reduce home heating and cooling energy consumption and lower associated greenhouse gas emissions.

1. Introduction

In 2012, the aggregate home energy expenditure of the 130 million dwellings in the USA [1] reached 10 Quads (a Quad is approx. 2.9×10^{11} kWh) [2]. This inordinate energy use stems from a diverse set of end-use activities, which includes space heating, ventilation and air conditioning (HVAC), water heating, cooking and lighting among others [3,4]. Across different climate zones, HVAC usage constitutes 59% of the northeast, 45% of the south, 59% of the midwest and 43% of the west USA building energy consumption [5,6]. This translates to a national average of 11 000 kWh spent on space conditioning per household [2], a notable portion of which is wasted due to inefficiencies [7,8]. With over 81% of the US population concentrated in urban areas [9], the state and federal governments have embraced important initiatives to reduce this waste and the associated carbon footprints of

cities by providing stimulus funds to adopt energy efficiency programmes [10,11]. These programmes, however, operate with limited resources in terms of tax rebates and technical assistance, and therefore can only support a selected number of buildings per year. These limitations call for fast and accurate methods to inform smart, citywide weatherproofing plans that pinpoint buildings with the greatest saving potential to minimize associated carbon emissions.

From an analytical perspective, there are two major challenges to construct such methods. One challenge is to develop reliable methodologies and toolsets to estimate energy consumption in buildings. There has been a vast body of literature on this subject since the 1970s that focuses on combining fundamental laws of thermodynamics with convective, conductive and radiative heat transfer to provide a consistent framework for buildings' energy modelling. Readers are referred to Swan & Ugursal [12], Kavgic *et al*. [13] and Zhao & Magoules [14], and references therein, for a comprehensive review of various complex dynamic modelling techniques. Another important challenge is concerned with developing a quantitative framework for assessment of energy-saving potentials that supports informed decisions on energy-saving policies relevant to retrofitting at an urban scale. Such a framework requires a scalable and sufficiently representative static description of a building's energy, i.e. gas and electricity, consumption that avoids redundant details while carrying enough physical information at a building level to inform weatherproofing options. This paper is motivated by the latter challenge and particularly focuses on gas consumption for space heating purposes in cold climates. Current approaches to evaluate retrofits use rating or audit tools to help energy-saving investments. Bardhan *et al*. [15] present an updated review of current practices and methods. The level of complexity and the required information vary significantly from one method to the other. The general results to inform policy are presented in terms of scores, characterizing the relative efficiency of a house in the region, or the recommendation of actions and the estimate of their potential savings. Current tools are often best suited for building-wise assessment. Upscaling the results to the community level relies on considering a 'typical' or 'average' house as a building block. For instance, on the Home Energy Saver (HES) website designed by the Lawrence Berkeley National Laboratory, Berkeley, CA (http://homeenergysaver.lbl.gov/consumer/), cities are analysed at the zip code level within a sizeable resident population from each urban area. These models are used to estimate the annual energy consumption of standardized houses in cities, to provide upgrade recommendations and to perform cost–benefit analyses of each specific retrofit. Although these tools show promise in helping energy-saving investment by filling the informational gap to a great extent, there are limitations in their accuracy and scalability when used to inform retrofittability at a city scale. These limitations deter their application as an effective and robust decision support methodology at the urban scale. We herein propose a method to simplify the estimates and to relate gas consumption from relevant information at a building level to the size of national sustainability goals. We perform an analysis of variance (ANOVA) by combining data from gas bills, buildings' footprints and physical simulations to avoid statistical bias introduced by 'typical houses'. We use this information to construct a simple yet efficient physics-based description of heating energy demand based on a model reduction scheme that encloses the most relevant parameters for the observed consumption. The model provides a means to identify buildings with the greatest potential for improvement, and quantifies the aggregated gas savings.

Gas demand patterns of residential, commercial and industrial sectors are driven by human activities, public perception and decisions, and physical constraints. Prevailing urban energy consumption models use micro-simulations to predict future usage by emulating the behaviour of urban dwellers (agents) and converting their decisions to respective demands [16,17]. Thus, a building's gas usage becomes a function of the choice of technology (e.g. insulation conditions, or efficiency of the heating system), its utilization (e.g. choice of the internal temperature set point) and additional extrinsic variables such as weather and neighbourhood patterns [18]. Retrofitting a building's thermal efficiency primarily targets the choice of technology, namely insulation to enhance the thermal resistance of its envelope (walls, windows, doors, roof and floors) and to reduce losses due to heat conduction, Q_{cond}, and infiltration, Q_{inf}. Thus, in order to quantify the potential gas savings of the retrofit of actual dwellings at the city scale, we study the diverse set of variables affecting consumption to derive a general statistical pattern of consumption in the studied urban settlement. Our goal is to identify, based on data analysis, the most influential set of physical parameters that operate locally [19,20]. We focus our solution on gas consumption for heating purposes and not on other means of consumption.

2. Results and discussion

The first step in this methodology is to understand the sensitivity of gas consumption response of a city's building stock to changes in external conditions, here being temperature. To this end, we combine buildings' footprints and associated gas consumptions with weather records for the same geographical location and time periods. We use actual monthly gas metre readings recorded by a utility company, E, per parcel (single- or multiple-family housing unit; this was the highest data resolution available) in kilowatt-hours (kWh) across the entire city. Herein, we use a 3-year-long record (2007–2009) collected for billing purposes, capturing the consumption pattern of almost 6200 individual residential buildings in Cambridge, MA, USA. After matching this record with the buildings' footprints from a geographic information system (GIS) dataset, we anonymized sources by removing addresses, in accordance with our non-disclosure agreement (NDA) terms. We match these data with the mean monthly temperature calculated via averaging the hourly temperature records from the closest weather station, which is located at Logan International Airport [21]. While the heat island phenomenon is certainly important in proper estimation of outdoor temperature in urban areas, we neglect its effect as it is out of the scope of the present study. Our data assimilation indicates that the gas consumption exhibits a characteristic piecewise linear form with respect to the outdoor temperature (figure 1*a*), separated by a cut-off temperature, T_0. The gas consumption increases linearly below this outdoor cut-off temperature, and it does not vary significantly for higher temperatures; thus it defines a temperature-insensitive baseline gas consumption (E_0), which is most likely due to hot water production. We identify this cut-off temperature as the temperature below which

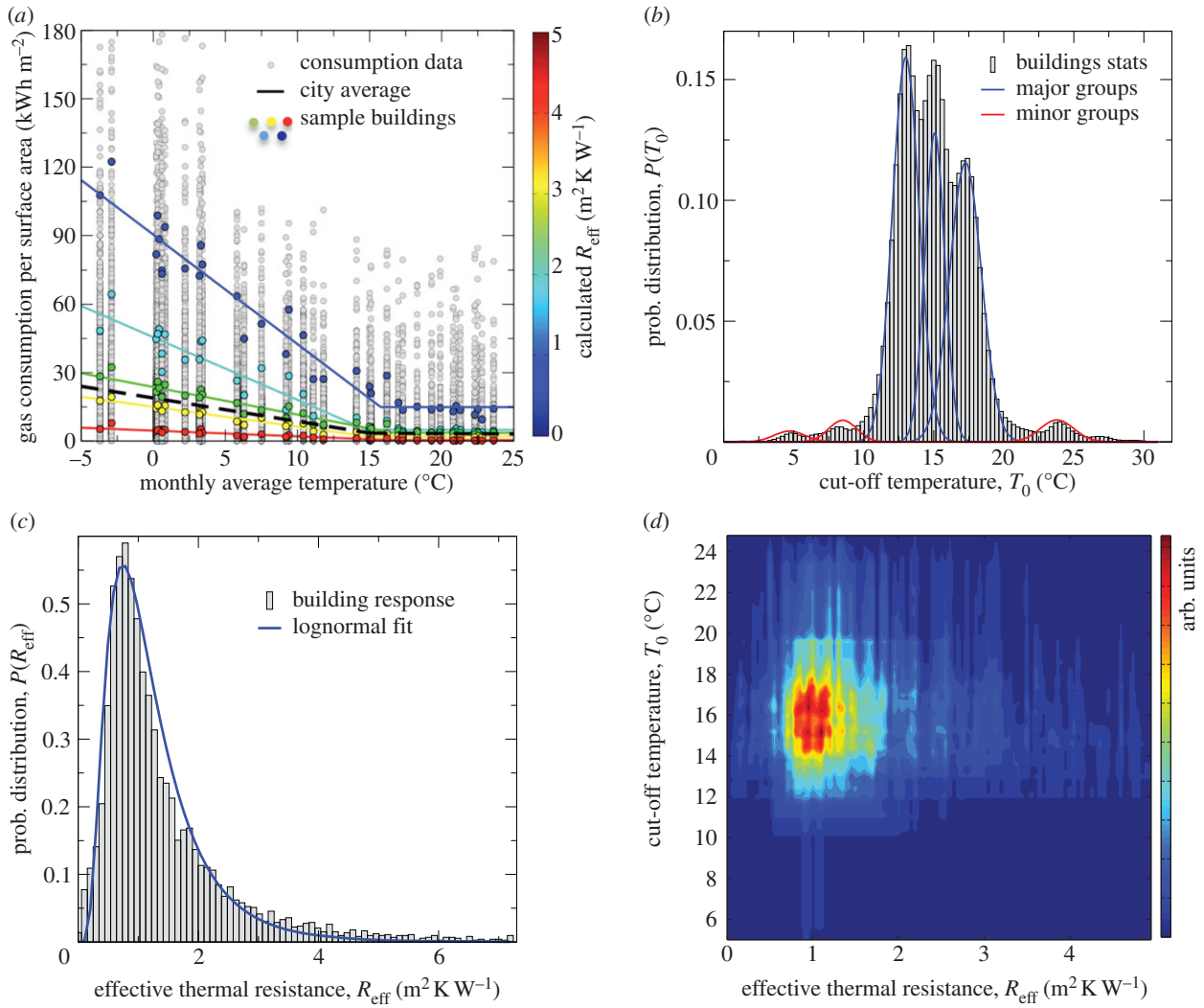

Figure 1. Data assimilation and analysis by integration of various data sources including buildings' footprints, weather data and gas consumption bills. (*a*) Gas consumption per surface area as a function of monthly average temperature for more than 6200 buildings in a period of 2007–2009 in Cambridge, MA, USA. The average city consumption per surface area is shown by the dashed black line. Five sample buildings are highlighted based on their effective thermal resistance defined in equation (2.1). The gas consumption of the buildings exhibits a piecewise linear trend with gas consumption increasing below a certain temperature threshold. (*b*) Distribution of the cut-off temperature indicating the variability of consumers' perception of and resistance to cold weather. (*c*) Distribution of effective thermal resistance for all the buildings. The Gaussian mixture analysis identifies the three major populations of the consumers that turn on their heating system at 17°C, 15°C and 13°C on average. The distribution of effective normal resistance follows a lognormal distribution, with an average value of 1 $m^2K\ W^{-1}$. (*d*) The joint probability distribution function of the T_0 and R_{eff}. The absence of correlation suggests that the energy consumption of buildings and the consumers' behaviour are not correlated.

consumers turn on their homes' heating systems to maintain indoor spaces at a desired comfort temperature, T_{comf}. In other words, the outdoor cut-off temperature is representative of individual choices of set point temperature. The probability density function of T_0 for the analysed building sample (figure 1*b*) sheds some light on these choices, in the form of three major peaks at 13°C, 15°C and 17°C (with 1°C s.d.), elucidating the core of the distribution (93% of the overall distribution, while the remaining 7% have medians at 4.7°C, 8.5°C and 23°C). These empirical findings give useful information to account for the distribution of the comfort level in buildings' energy simulations at the city scale.

For gas consumption, each building has a constant rate in the increase of gas consumption below the offset temperature, suggesting a linear form between the heating energy and outside temperature in excess of the baseline gas consumption (E_0), in the form:

$$E = \frac{1}{R_{eff}}(T_0 - T_{out}) \times t \times S, \qquad (2.1)$$

where S is the building's envelope surface area and E is the heating energy necessary to maintain an inside temperature of T_{comf} when the outside temperature T_{out} is below T_0 during the time interval of exposure, t, which corresponds to the numbers of hours between two consecutive energy readings (E) by the utility company. Moreover, the linearity between the temperature difference $T_0 - T_{out}$ and the heating energy defines a linear coefficient, R_{eff} (in $m^2K\ W^{-1}$), that can be viewed as an effective thermal resistance representative of the thermal efficiency of a building. Unlike the effective heat loss rate in the degree-day approaches [22–24], we expect R_{eff} to depend only on the physical attributes of a building's envelope; namely, heat transport, radiation and infiltration properties. For the sample of 6200 homes in Cambridge, MA, USA, the effective thermal resistance that is obtained by a linear fitting of the energy readings according to equation (2.1) is found to follow a lognormal distribution (figure 1*c*). This lognormal distribution stems from the multiplicative random processes influencing the effective thermal resistance. The fact that this distribution is uncorrelated

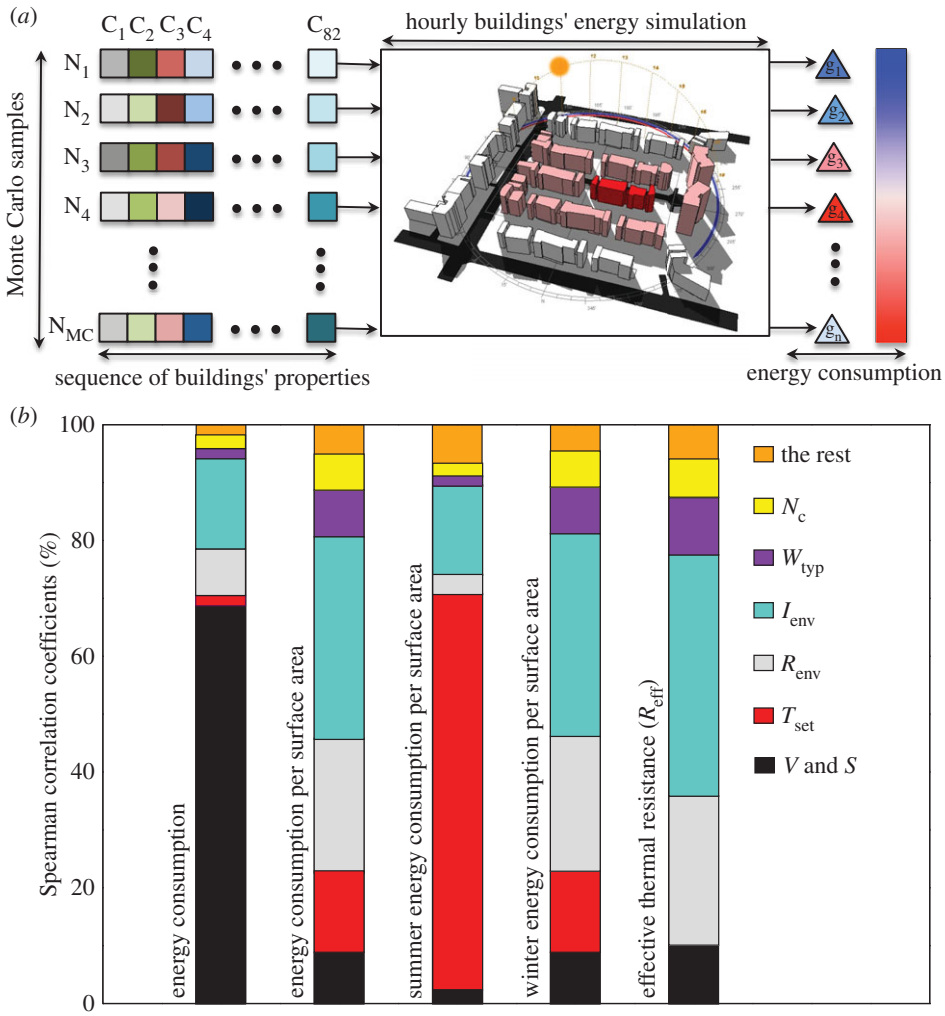

Figure 2. Reducing the complexity and identification of the most influential parameters in buildings' gas consumption via analysis of variance. (*a*) The schematic representation of the Monte Carlo uncertainty propagation method. Given the probabilistic nature of buildings' characteristics at the city scale denoted by 82 squares (C_1-C_{82}), the gas consumption can be viewed as an uncertain parameter. Here, building energy consumption modelling bridges the gap between each building's characteristic space and the gas consumption space. We randomly sample from the building's characteristic space and calculate the relevant energy consumption (denoted by triangles). (*b*) Identification of the most influential parameters from the GSA on different consumption norms. While the gas consumption is strongly affected by building volume and surface area (V, S), the gas consumption per surface area contains more information about the air infiltration rate (I_{env}), thermal resistance of walls (R_{env}) and window type (W_{typ}). Unlike winter gas consumption, the summer consumption depends strongly on the internal temperature set point (T_{set}) due to the proximity of the outdoor and indoor temperatures. More importantly, T_{set} does not affect R_{eff} because if T_{set} is relatively constant inside a building then the temperature gradient is only due to the variation of the outdoor temperature.

with T_0 (figure 1*d*) is potentially related to the so-called rebound [25,26] or takeback effect [27], i.e. occupants' tendency to forgo taking advantage of high R_{eff} by increasing the indoor temperature. Also, this uncorrelated observation can possibly stem from the fact that the energy costs might have been included in dwellings' monthly rents, hence directly affecting the tenants' choice in adopting conservative behaviour. There are many other facets to occupants' choice, especially with regards to gas consumption. To infer these aspects along with HVAC performance metrics and their dependence on external conditions (temperature, humidity, etc.), we require high-resolution hourly consumption data that are currently only available at the building level [28–31].

At this stage, to identify individual buildings with the highest retrofitting potential from billing information alone, it suffices to link R_{eff} with the physical parameters that affect buildings' gas consumption in their specific environment. To establish this link, we resort to buildings' energy consumption modelling (see Material and methods section). We create a probabilistic gas consumption model of a block of nine buildings that interact with each other via shadowing and thermal interactions when they are physically in contact. This approximation neglects the shadowing effect of far-field tall buildings. We subsequently propagate the uncertainty in a set of 82 input parameters that affect buildings' energy consumption using the Energyplus package [32] (figure 2*a*; see the electronic supplementary material, table S1, for the entire list of uncertain input parameters, their distribution type and ranges of variations). Here, we employ global sensitivity analysis (GSA) to shed light on the relative importance of individual factors affecting heating energy consumption [33–37] (see Material and methods section).

Note that, for the purpose of this approach, we fixed the efficiency of the simplified heating system efficiency, η_H, in our Energyplus simulations. This parameter is a critical one, and is commonly related to the resulting scaling in buildings' gas consumption. We do so, because our goal here is to simplify the role of the physical parameters in the

resulting consumption for heating purposes. In practice, one would need to scale the resulting expression, which is based on physical parameters, by introducing the exact value of η_H of the building under consideration.

After performing GSA on Energyplus simulation results, we find that only seven building parameters are important in determining heating energy consumption at the monthly level (figure 2*b*). These seven variables are the: building's volume and envelope surface area (V and S), number of neighbours sharing a wall with the building (N_c), effective thermal resistance of the building envelope (R_{env}), air infiltration rate (I_{env}), average temperature set point (T_{set}) and window type indicating the number of glazing (W_{typ}). In future work, one could potentially combine N_c and S into a single variable such as exposed surface area. As a key result, the complexity of the problem is now reduced to its most influential variables; however, this form still presents an opportunity for further simplification.

Based on the sensitivity analysis presented in figure 2*b* performed via systematic simulations of Energyplus and the exploration of all the parameter space set in its relevant ranges, we note that, while the largest contribution to the variance of gas consumption (E) is accounted for through the variability in building size, the Spearman rank correlation coefficient (SRCC) in gas consumption per surface area (E/S) is mainly the result of the interplay between individual activity and envelope properties (T_{set}, I_{env}, R_{env}). Therefore, E/S is more informative about the thermal efficiency of the building envelope *per se*. Also, most of the contribution to the variance of E^{summer}/S is attributable to the consumer's set point temperature, T_{set}. This is likely to be the result of the fluctuating temperatures of the summers in northeastern USA. In other words, the average monthly temperature is close to T_{comf}, making the temperature difference between the inside and outside sensitive to individual choices and not the building's thermal efficiency. However, the average monthly temperature in cold seasons drops significantly and steadily below T_{comf}, which, as shown in figure 2*b*, results in gas consumption (E^{winter}) that can be expressed in terms of building variables (I_{env}, R_{env}) and different choices per building (T_{set}). When inspecting the ANOVA results of R_{eff}, we notice that R_{eff} has the same characteristics as E^{winter}/S with the exception of being completely independent of residents' choices (T_{set}). This reflects the preferences of households to maintain T_{set} at the monthly level regardless of T_{out}. As a consequence, the gradient of energy consumption as a function of T_{out} is independent of T_{set} and captures solely the building efficiency properties. Thus, consistent with the findings from the energy data analysis for Cambridge, MA, USA, we confirm from ANOVA and buildings' energy simulations that R_{eff} is the most convenient norm of choice for capturing the physical response of buildings, reducing the consumption to only physical variables in buildings.

Our task is then reduced to quantitatively describe R_{eff} as a function of physical properties of each building and to quantify the impact of weatherproofing at the city scale. This is achieved here by simplifying the problem through a dimensional analysis of the physical quantities involved that possibly affect R_{eff}, namely the effective thermal resistance of the building envelope (R_{env}), the air infiltration rate (I_{env}), the volumetric heat capacity of air (C_v^{air}) emphasizing that the heat exchange is performed through air, the building's characteristic dimension expressed by the volume-to-surface area ratio (V/S) and the efficiency of the HVAC system (η_H). This analysis allows us to further reduce the dimension of the problem to a three-parameter relation between the dimensionless thermal resistance, the ratio of infiltration (Q_{inf}) to conduction losses (Q_{cond}) and the thermal efficiency of the HVAC system (see Material and methods and the electronic supplementary material, section V, for detail on the dimensional analysis):

$$\Pi_1 = \frac{R_{eff}}{R_{env}} = F\left(\Pi_2 = R_{env}.I_{env}.C_v^{air}\frac{S}{V} = \frac{Q_{inf}}{Q_{cond}}, \Pi_3 = \eta_H\right). \quad (2.2)$$

This implies that a simple functional form of $\Pi_1 = F(\Pi_2, \Pi_3)$ is sufficient to describe the physical response of the system without the need to run Energyplus simulations repeatedly. To determine this functional relation, a full factorial design in the (R_{env}, I_{env}, V/S) space is performed by means of Energyplus simulations (figure 3*a*). By enforcing the law of conservation of energy, the response of the simulations can be written as (see the electronic supplementary material, section V):

$$\Pi_1 = \frac{R_{eff}}{R_{env}} = \frac{1}{\Pi_3 \times (A_1 + A_2 \times \Pi_2)}, \quad (2.3)$$

where A_1 and A_2 are the degrees of freedom in the model. While the functional form of the dimensionless relation in equation (2.3) remains unaltered irrespective of climate, A_1 and A_2 are strongly dependent on the location and climate under consideration. In practice, these parameters need to be calibrated with results of physical simulations in the urban settlement under a given climate. Also, these parameters need to be calibrated with results of physical simulations in the urban settlement under consideration (for instance, for the case of detached buildings with double-glazed windows, $A_1 = 0.49$ and $A_2 = 0.30$; figure 3*a*). The dimensionless form in equation (2.3) provides insights into a building's thermal efficiency from the perspective of a simplified complex system. Unlike the effective heat loss rate in the PRISM approach [38], R_{eff} and its dimensionless functional form not only account for weather normalization but also provide a quantitative framework to compare dwellings with different sizes. Nonetheless, the individual physical properties of buildings cannot be uniquely identified using the dimensionless model in equation (2.3). For instance, as shown in figure 3*b*, higher thermal efficiency at the building level can be equally achieved by increasing the thermal resistance of the envelope or by decreasing the air infiltration rate. That is, all weatherproofing solutions are located on an iso-performance line [39] in the (R_{env}, I_{env}) space at a fixed η_H, where any point corresponds to a unique value of R_{eff}. This finding was further tested via standard machine learning methods, namely by means of multiple adaptive regression splines (MARS) [40,41], which are well suited for capturing response surfaces of multi-parametric problems [42,43]. These results demonstrate that, given the nature of the gas consumption response to the parameter space, the response of the system as a surrogate function is always solvable.

In general, equation (2.3) is a powerful tool for simplifying decisions and first-order estimates of energy savings in a sense that, if we know W_{typ}, R_{env}, I_{env} and η_H of a given building prior to the retrofit, we will have a robust estimation

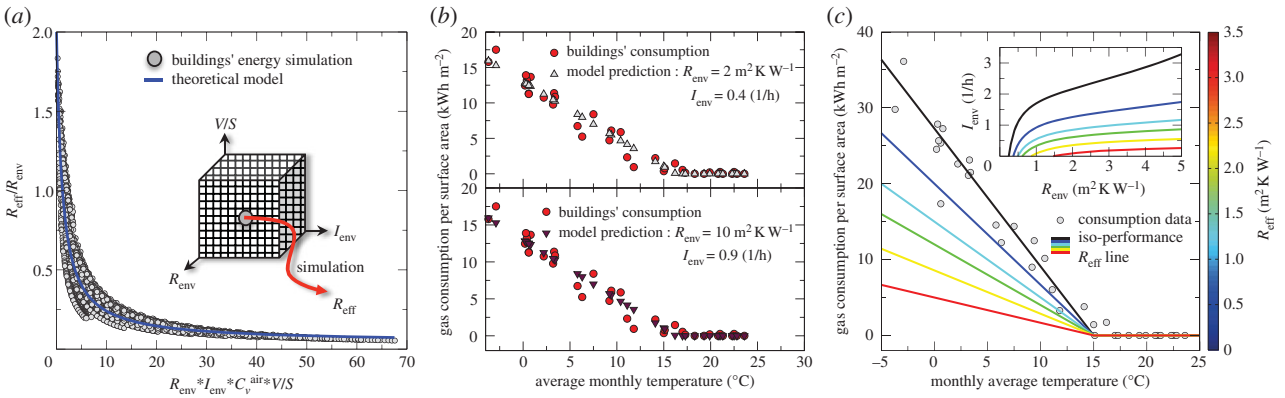

Figure 3. Estimation of buildings' gas consumption in dimensionless space via the response surface area methodology. (*a*) Construction of the surrogate model in dimensionless form. Considering a dense full factorial grid in (I_{env}, R_{env}, *V*/*S*) space (as denoted in the inset), R_{eff} can be calculated at each point via building energy simulation using Energyplus. For a building with a given W_{typ} and N_o the dimensional analysis yields that there exist only three dimensionless quantities relating R_{eff} to the rest of the influential parameters, I_{env}, R_{env}, *V*/*S*, C_v^{air} (volume heat capacity of air) and η_H, which are: $\pi_1 = R_{eff}/R_{env}$, $\pi_2 = I_{env} \times R_{env} \times C_v^{air} * V/S$ and $\pi_3 = \eta_H$. A surrogate model of the form $\pi_1 = 1/\pi_3(C_1 + C_2 \times \pi_2)$, as shown by the blue line, is fitted to the simulation results. Having the surrogate model at hand, we can estimate the performance of buildings without actually running the computationally expensive simulations. (*b*) Identification of buildings' thermal properties from monthly gas consumption data using the surrogate model. The solution of inverse problems to identify the buildings' characteristics such as I_{env} and R_{env} is non-unique. For instance, the surrogate model predicts the same R_{eff} value at two alternative cases of ($I_{env} = 0.4$ (1/h), $R_{env} = 2$ m^{2}K W^{-1}) and ($I_{env} = 0.9$ (1/h), $R_{env} = 10$ m^{2}K W^{-1}) located on an iso-performance line. (*c*) Estimation of retrofit gas savings at the building level. Given the monthly gas consumption of the building, R_{eff} is estimated by fitting a line denoted in black. The associated R_{eff} attributes to the black iso-performance line in the inset in (R_{env}, I_{env}) space. By performing retrofit, the R_{eff} of the building increases. The colours of iso-performance lines match the colours for gas consumption per surface area in the main graph. Therefore, the higher increase in R_{eff} directly translates to higher gas savings upon retrofit.

of the energy-saving potential for each retrofit scenario. In fact, we can obtain the best cost-effective retrofit scenario by juxtaposing this reduced order model with collected gas consumption records and buildings' characteristics measured via on-site home inspection. Although we have effectively reduced the number of influential parameters, W_{typ}, R_{env}, I_{env} and η_H of the majority of buildings are not available at the urban level. Therefore, we shift our focus to the collective response of these variables, which is manifested in R_{eff}. Here, we start at the building level by considering the gas consumption of an arbitrary building and its pertinent linear fit (figure 3*c*). The obtained effective heat resistance prior to retrofit, R_{eff}^-, is situated on an iso-performance line (black solid line in the inset of figure 3*c*), which captures the current thermal performance of the building in terms of envelope heat resistance (R_{env}), infiltration rate (I_{env}) and HVAS efficiency (η_H) according to equation (2.3). Any weatherproofing option, such as increasing insulation (R_{env}), reducing air infiltration rate (I_{env}), improving the efficiency of the heating system (η_H) or installing multiple-paned windows (W_{typ}), while retaining the preferred behavioural choices (neglecting the rebound effect [25–27] by assuming the same value of T_0 in figure 3*c*) would entail an increase of $R_{eff}^- \rightarrow R_{eff}^+$, to higher iso-performance levels (smaller slope of energy consumption in figure 3*c*), and thus, in light of equation (2.1), to an energy saving after retrofitting of $\Delta E = E^- - E^+ = (1 - (R_{eff}^-/R_{eff}^+))(E^- - E_0)$, which is independent of the particular choice of weatherproofing.

This simple form provides a straightforward means to upscale the gas-saving potential from the building to the city scale to assist with science-informed urban policy choice and implementation. That is, the challenge pertaining to city-scale strategic retrofit planning is concerned with finding the shortest path to retrofit that achieves the highest savings with the least number of retrofitted buildings. In this regard, an important feature emerges from the ranking of the potential gas saving of buildings calculated in the same manner as presented in figure 3*a*. We find that the rank and magnitude of gas savings follow over a large range a power law with an exponent of 0.75; much like Zipf's law [44,45]. While the deviation of the tail from Zipf's law is attributed to buildings with insignificant gas savings, the tail of the gas-saving distribution follows a power law with an exponent of 2.2 (inset of figure 4*a*). Given the significance of a Zipf-type data distribution, it appears to us that such a ranking based on gas-saving potential will provide the shortest path for city-scale gas savings. To test our hypothesis, we compare this ranking with other selection criteria associated with urban policy choices, starting with a random retrofit of buildings at the city scale, performed upon requests of building owners, in which case the achieved gas saving scales linearly with the number of retrofits. The results of this analysis, displayed in figure 4*b*, show that an informed selection based on ranking the energy saving of buildings (rank(ΔE)) provides indeed the highest rate of energy saving per retrofit, followed by an informed selection based on ranking of buildings' gas consumption-per-surface area (rank($E^- - E_0/S$)), building sizes (rank(S, V)) and effective thermal resistance (rank(R_{eff}^-)). When targeting buildings with high priority and after on-site inspections, equation (2.3) can quantitatively predict potential gas savings for various retrofit scenarios. By way of example, if Cambridge, MA, USA, targets a 40% overall gas consumption reduction related to heating, it would suffice, with such an informed selection process, to retrofit only 16% of the entire building stock as mapped in figure 4*c* in order to achieve this goal, in contrast to 67% of buildings with a random selection procedure to achieve the same target by neglecting the rebound effect [25–27]. That is, the proposed selection scheme based on ranking potential

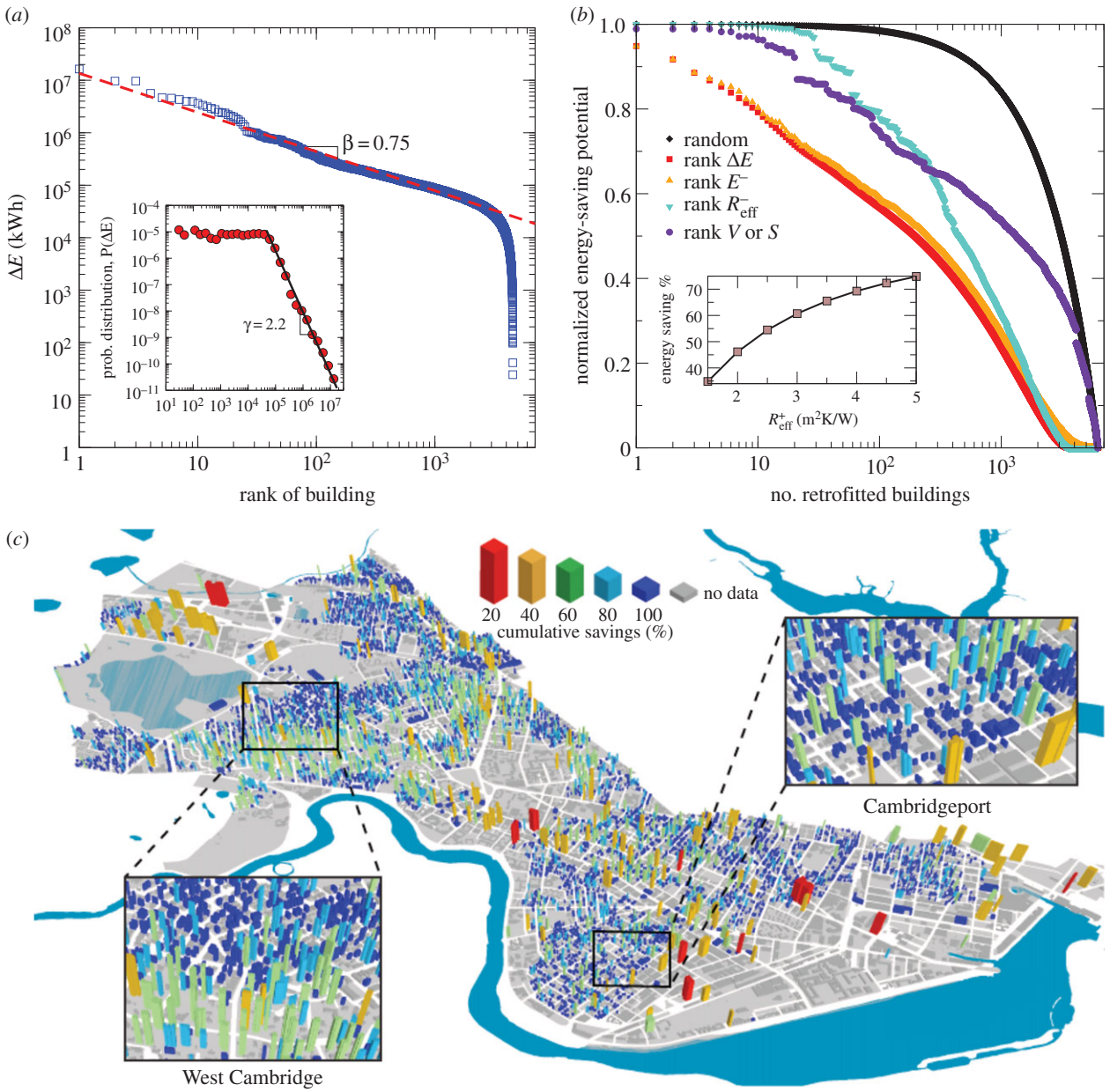

Figure 4. Citywide retrofittability analysis by combining GIS, weather, gas consumption data and surrogate modelling at the individual building level. (*a*) Gas saving as a function of the retrofit rank of a building, which follows a Zipf-like relation with the exponent of $\beta = 0.75$. The inset displays the distribution of ΔE, which indicates a linear tail in log–log scale akin to scaling laws. (*b*) Comparative study between different retrofit scenarios based on ranking of ΔE, E^{-}, R_{eff}^{-}, V or S and demonstrating their effectiveness against a random scenario (blind retrofit) at the city scale. The calculation of gas-saving potential at the city level is based on an assumption of $R_{\text{eff}}^{+} = 3.0\,\text{m}^2\text{K}\,\text{W}^{-1}$ after retrofit. If buildings are retrofitted on a random basis, the gas consumption of the city decreases linearly with the number of retrofits. If the buildings are retrofitted based on their potential gas saving, the city-scale gas consumption decreases nonlinearly through the shortest path strategy. The GIS map of buildings in Cambridge, MA, USA, with colours representing the retrofittability potential at the city level. The colours are randomized to respect the privacy laws and terms of NDA required by the utility company.

gas savings provides an efficient means to achieve the shortest path for substantial energy savings at the city scale.

To conclude, we propose a method of analysis that combines data on gas consumption, climate and buildings' footprints with surrogate energy modelling. This powerful framework reduces the complexity of the problem to a simple functional form to estimate the thermal response of buildings. Calibrated with utility data, this functional form allows us to easily estimate potential gas savings per building under different retrofit scenarios with minimal computational expense. When applied at the urban scale, we can make informed selections towards the reduction of the gas consumption footprint by identifying the shortest path to the desired goal. The method is portable to cities in different climates, requiring solely data that are readily available for billing and urban planning purposes. This physical approach would benefit from the interaction with new advancement in materials design [46,47] and policy analyses to shed light on energy price-dependent [48], cost-effectiveness [49] and properties' tenure-dependent considerations [50] in various city-scale retrofit scenarios. Hence, from a practical view point, we consider mitigation of city-scale gas consumption and associated carbon emissions to be a multi-objective optimization problem characterized by a Pareto front in the space of technical, economical, legal and political aspects. Similar model reduction approaches combining large data with statistical analysis and physical simulations to gain predictive understanding of the system's response appear to us

to be promising for urban energy solutions such as patterns of hourly electric demand and the adoption of alternative sources for generation of electricity. These methods have the premise to help cities to use pervasive data sources to optimize decisions that make them more environmentally and economically sustainable.

3. Material and methods

3.1. Estimation of T_0 and R_{eff} via a piecewise linear regression

The correlation between monthly heating gas consumption and monthly average outdoor temperature exhibits a piecewise linear trend. Such a gas consumption trend, Y, is mathematically expressed as:

$$\Upsilon(E_0, T_0, R_{\text{eff}}) = \frac{E_0}{S} - \frac{1}{R_{\text{eff}}} \times H(T_0 - T) \times (T - T_0), \qquad (3.1)$$

where $H(x - x_0)$ is the Heaviside step function with x_0 being the step position. To find the best piecewise linear regression, or, in other words, the best $(E_0, T_0, R_{\text{eff}})$ triplet for a building, we minimize the L_2 norm of regression error, defined as:

$$L_2(E_0, T_0, R_{\text{eff}}) = \left| \Upsilon - \frac{E}{S} \right|^2, \qquad (3.2)$$

where E/S is the actual heating energy consumption per surface area. We performed all regression steps for the entire dataset in an automatic fashion with no data manipulation or treatment, as this makes the analysis rather subjective. The electronic supplementary material, section VI, includes a Matlab script developed for piecewise linear regression. Figure S4 in the electronic supplementary material shows the distribution of the regression coefficient of determination, R^2, indicating that the majority of buildings in Cambridge, MA, USA, follow the aforementioned piecewise linear trend.

3.2. Buildings' energy consumption modelling

We performed building energy simulations using the standard Energyplus package [32]. For this purpose, we constructed an hourly weather file for the period of 2007–2009 using the weather measurements recorded at Logan International Airport by the National Oceanic and Atmospheric Administration [21] (see the electronic supplementary material, section I, for details). We subsequently performed a local sensitivity analysis to uncover the extent of building interactions via the shadowing effect (see the electronic supplementary material, section III and figure S2). This analysis shows that only the first eight neighbours affect a building's heating energy consumption. Hence, we performed further simulations by only considering a block of nine buildings (see the electronic supplementary material, figure S3) that interact through the shadowing effect by considering the Sun's path in the sky dome. The hourly energy consumption predictions are extremely fine compared with actual energy measurements, i.e. monthly energy bills. This suggests that energy simulation and predictions should be relevant to average monthly trends rather than spontaneous temporal variations. Therefore, we considered an average occupancy level and constant indoor temperature in our simulations and aggregated hourly energy predictions to monthly values. Afterwards, we constructed a probabilistic model of buildings' energy consumption by considering several uncertain parameters (see the electronic supplementary material, section IV and table S1). In particular, to be representative of a city's texture, the distances between buildings in the simulations are taken from the distribution of inter-building distances calculated from analysis of the GIS dataset (see the electronic supplementary material, section II and figure S1).

3.3. Complexity reduction via global sensitivity analysis

To reduce the parameter space, we performed a sensitivity analysis using Monte Carlo sampling to propagate the uncertainty from all parameters containing all possible building sizes and specifications (figure 2*a*) into energy consumption space. This Monte Carlo sampling provides a probabilistic mapping necessary to infer the contribution of each uncertain variable to the variance of energy consumption using ANOVA. In particular, we employed the SRCC to characterize the sensitivity of energy consumption norms with respect to all uncertain variables (see the electronic supplementary material, section IV).

3.4. Dimensional analysis and response surface modelling

From a dimensional perspective, the effective thermal resistance of an envelope with dominant conduction and infiltration heat transfer mechanisms can be written as $R_{\text{eff}} = \Psi(R_{\text{env}}, I_{\text{env}}, V, S, C_{\text{v}}^{\text{air}}, \eta_{\text{H}})$. The dimensional analysis reduces this functional form to a simple relation between fewer numbers of dimensionless parameters. The rank of the exponent matrix, the matrix formed by the exponents of the variables' dimensions, is 4 (see the electronic supplementary material, section V). Thus, according to the π-theorem [51], there are only two independent dimensionless variables among the initial six parameters. The dimensionless relation is: $\Pi_1 = R_{\text{eff}}/R_{\text{env}} = F(\Pi_2 = R_{\text{env}}.I_{\text{env}}.C_{\text{v}}^{\text{air}}(V/S), \Pi_3 = \eta_{\text{H}})$, where Π_1 is the ratio of the effective thermal resistance of the system to the conduction resistance of the envelope and Π_2 will be shown to be the ratio of the infiltration heat transfer to the conductive heat transfer. Dimensional analysis effectively reduces the number of variables but it does not quantify the relation between them. To this end, we used the conservation of energy law to propose a functional form of F. The conservation of energy for the control volume (volume inside an envelope), which exchanges heat with surrounding media through conduction and infiltration, can be written as

$$Q_{\text{tot}} = A_1 \frac{S.\Delta T}{R_{\text{env}}} + A_2.I_{\text{env}}.C_{\text{v}}^{\text{air}}.V.\Delta T, \qquad (3.3)$$

where Q_{tot} is the total heat loss through the envelope and thus is equal to $S.\Delta T/R_{\text{eff}\times\eta_{\text{H}}}$, which can be rearranged in the form of equation (2.3). Coefficients A_1 and A_2 can be identified via either simulation or experiment. Here, we used Energyplus software to numerically estimate these coefficients. We have performed a full factorial simulation varying R_{env}, I_{env} and V/S at a fixed efficiency of the HVAC system (see the electronic supplementary material, section V). R_{eff} is computed as the derivative of predicted energy consumption with respect to average monthly temperature. The results are plotted in the $\Pi_1 - \Pi_2$ space and A_1 and A_2 are derived by fitting equation (2.3) to the results via the least-squares approach.

Authors' contributions. M.C.G., F.-J.U., R.J.-M.P., M.J.A.Q. and J.F. designed the project. M.J.A.Q., J.M.S. and J.T. performed the energy, GIS and weather data assimilation. M.J.A.Q. performed the Energyplus simulations. M.J.A.Q. and A.N. performed the sensitivity analysis. M.J.A.Q., A.N., M.C.G. and F.-J.U. performed the dimensional analysis, designed the reduced order model and interpreted the potential energy savings based on the surrogate model. All authors contributed to writing the manuscript.

Funding. Partial financial support through the Concrete Sustainability Hub at MIT with sponsorship provided by the Portland Cement Association and the Ready Mixed Concrete Research & Education Foundation is also acknowledged. R.J.-M.P. and F.-J.U. wish to acknowledge the support of the ICoME2 Labex (ANR-11-LABX-0053) and the A*MIDEX projects (ANR-11-IDEX-0001-02), cofounded by the French programme 'Investissements d'Avenir' managed by the ANR, the French National Research Agency.

Competing interests. We declare we have no competing interests.

Acknowledgements. M.J.A.Q. acknowledges discussions with S. Do and K. Goldstein. M.C.G. acknowledges the support of the MIT-Accenture alliance and Center for Complex Engineering Systems (CCES) at KACST, and J.T. acknowledges an NSF graduate studies fellowship. M.J.A.Q. acknowledges partial funding from the Henry Samueli School of Engineering, University of California Irvine.

References

1. American Housing Survey for the United States. 2007 U.S. Census Bureau, Current Housing Reports, Series H150/07., 2008. See http://www.census.gov/prod/2008pubs/h150-07.pdf.
2. U.S. Environmental Protection Agency. 2009 Buildings and their impact on the environment: a statistical summary. Green Building Workshop, U.S. Environmental Protection Agency (revised 22 April, 2009). See http://www.epa.gov/greenbuilding/pubs/gbstats.pdf.
3. Shimoda Y, Asahi T, Taniguchi A, Mizuno M. 2007 Evaluation of city-scale impact of residential energy conservation measures using the detailed end-use simulation model. *Energy* **32**, 1617–1633. (doi:10.1016/j.energy.2007.01.007)
4. Sartori I, Hestnes AG. 2007 Energy use in the life cycle of conventional and low-energy buildings: a review article. *Energy Build*. **39**, 249–257. (doi:10.1016/j.enbuild.2006.07.001)
5. Berry J. 2009 *Residential energy consumption survey (RECS) data*. U.S. Energy Information Administration. See http://www.eia.gov/consumption/residential/data/2009/.
6. Center for Climate and Energy Solutions. 2013 *Leveraging natural gas to reduce greenhouse gas emissions*. See http://www.c2es.org/publications/leveraging-natural-gas-reduce-greenhouse-gas-emissions.
7. American Physical Society. 2008 *Energy future: think efficiency report*. See http://www.aps.org/energyefficiencyreport/report/index.cfm.
8. Lawrence Livermore National Laboratory. See https://flowcharts.llnl.gov/.
9. United Nation's Department of Economics and Social Affairs. 2015. See http://esa.un.org/unpd/wup/.
10. The White House. 2013 *The President's climate action plan*. See http://www.whitehouse.gov/sites/default/files/image/president27sclimate actionplan.pdf.
11. Mass Save. 2015 Energy saving programs. See http://www.masssave.com/.
12. Swan LG, Ugursal VI. 2009 Modeling of end-use energy consumption in the residential sector: a review of modeling techniques. *Renew. Sustain. Energy Rev*. **13**, 1819–1835. (doi:10.1016/j.rser.2008.09.033)
13. Kavgic M *et al*. 2010 A review of bottom-up building stock models for energy consumption in the residential sector. *Build. Environ*. **45**, 1683–1697. (doi:10.1016/j.buildenv.2010.01.021)
14. Zhao H, Magoules F. 2012 A review on the prediction of building energy consumption. *Renew. Sustain. Energy Rev*. **16**, 3586–3592. (doi:10.1016/j.rser.2012.02.049)
15. Bardhan A, Jaffee D, Kroll C, Wallace N. 2014 Energy efficiency retrofits for U.S. housing: removing the bottlenecks. *Reg. Sci. Urban Econ*. **47**, 45–60. (doi:10.1016/j.regsciurbeco.2013.09.001)
16. Ben-Akiva ME, Lerman SR. 1985 *Discrete choice analysis: theory and application to travel demand*. Cambridge, MA: MIT Press.
17. Chingcuanco F, Miller EJ. 2012 A microsimulation model of urban energy use: modelling residential space heating demand in ILUTE. *Comput. Environ. Urban Syst*. **36**, 186–194. (doi:10.1016/j.compenvurbsys.2011.11.005)
18. Reinhart CF, Herkel S. 2000 The simulation of annual daylight illuminance distributions—a state-of-the-art comparison of six RADIANCE-based methods. *Energy Build*. **32**, 167–187. (doi:10.1016/S0378-7788(00)00042-6)
19. Bettencourt L, West G. 2010 A unified theory of urban living. *Nature* **467**, 912–913. (doi:10.1038/467912a)
20. Kolter JZ, Ferreira Jr J. 2011 A large-scale study on predicting and contextualizing building energy usage. In *Proc. Twenty-Fifth AAAI Conf. Artif. Intell., San Francisco, CA, 7–11 August 2011*. See http://dspace.mit.edu/handle/1721.1/77192.
21. National Oceanic and Atmospheric Administration. 2014 See http://www.noaa.gov/.
22. Quayle RG, Diaz HF. 1980 Heating degree day data applied to residential heating energy consumption. *J. Appl. Meteorol*. **19**, 241–246. (doi:10.1175/1520-0450(1980)019<0241:HDDDAT>2.0.CO;2)
23. Sailor DJ, Munoz JR. 1997 Sensitivity of electricity and natural gas consumption to climate in the U.S.A.—methodology and results for eight states. *Energy* **22**, 987–998. (doi:10.1016/S0360-5442(97)00034-0)
24. Manfren M, Aste N, Moshksar R. 2013 Calibration and uncertainty analysis for computer models—a meta-model based approach for integrated building energy simulation. *Appl. Energy* **103**, 627–641. (doi:10.1016/j.apenergy.2012.10.031)
25. Berkhout PHG, Muskens JC, Velthuijsen WJ. 2000 Defining the rebound effect. *Energy Policy* **28**, 425–432. (doi:10.1016/S0301-4215(00)00022-7)
26. Haas R, Biermayr P. 2000 The rebound effect for space heating. Empirical evidence from Austria. *Energy Policy* **28**, 403–410. (doi:10.1016/S0301-4215(00)00023-9)
27. Goldberg M, Fels M. 1986 Refraction of prism results into components of saved energy. *Energy Build*. **9**, 169–180. (doi:10.1016/0378-7788(86) 90018-6)
28. Kreider J *et al*. 1995 Building energy use prediction and system-identification using recurrent neural networks. *J. Sol. Energy Eng.-Trans. Asme* **117**, 161–166. (doi:10.1115/1.2847757)
29. Dhar A, Reddy TA, Claridge DE. 1999 A Fourier series model to predict hourly heating and cooling energy use in commercial buildings with outdoor temperature as the only weather variable. *J. Sol. Energy Eng*. **121**, 47–53. (doi:10.1115/1.2888142)
30. Li Q, Meng Q, Cai J, Yoshino H, Mochida A. 2009 Applying support vector machine to predict hourly cooling load in the building. *Appl. Energy* **86**, 2249–2256. (doi:10.1016/j.apenergy.2008.11.035)
31. Gonzalez PA, Zamarreno JA. 2005 Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. *Energy Build*. **37**, 595–601. (doi:10.1016/j.enbuild.2004.09.006)
32. Crawley DB *et al*. 2001 EnergyPlus: creating a new-generation building energy simulation program. *Energy Build*. **33**, 319–331. (doi:10.1016/S0378-7788(00)00114-6)
33. Tian W. 2013 A review of sensitivity analysis methods in building energy analysis. *Renew. Sustain. Energy Rev*. **20**, 411–419. (doi:10.1016/j.rser.2012.12.014)
34. Hopfe CJ, Hensen JLM. 2011 Uncertainty analysis in building performance simulation for design support. *Energy Build*. **43**, 2798–2805. (doi:10.1016/j.enbuild.2011.06.034)
35. Wilde P, Tian W. 2009 Identification of key factors for uncertainty in the prediction of the thermal performance of an office building under climate change. *Build. Simul*. **2**, 157–174. (doi:10.1007/s12273-009-9116-1)
36. Eisenhower B, O'Neill Z, Fonoberov VA, Mezić I. 2012 Uncertainty and sensitivity decomposition of building energy models. *J. Build. Perform. Simul*. **5**, 171–184. (doi:10.1080/19401493.2010.549964)
37. Mara TA, Tarantola S. 2008 Application of global sensitivity analysis of model output to building thermal simulations. *Build. Simul*. **1**, 290–302. (doi:10.1007/s12273-008-8129-5)
38. Fels M. 1986 Prism—an introduction. *Energy Build*. **9**, 5–18. (doi:10.1016/0378-7788(86)90003-4)
39. De Weck OL, Jones MB. 2006 Isoperformance: analysis and design of complex systems with desired outcomes. *Syst. Eng*. **9**, 45–61. (doi:10.1002/sys.20043)
40. Friedman JH. 1991 Multivariate adaptive regression splines. *Ann. Stat*. **19**, 1–67. (doi:10.1214/aos/1176347963)
41. Abdolhosseine Qomi MJ. 2014 From atoms to cities: a bottom-up analysis of infrastructure

materials and systems. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA. See http://hdl.handle.net/1721.1/99617.

42. Russell SJ, Norvig P, Davis E. 2010 *Artificial intelligence: a modern approach*. Englewood Cliffs, NJ: Prentice Hall.
43. Michalski RS, Michalski RS, Carbonell JG, Mitchell TM. 1986 *Machine learning: an artificial intelligence approach*. Los Altos, CA: Morgan Kaufmann.
44. Gabaix X. 1999 Zipf's law for cities: an explanation. *Q. J. Econ*. **114**, 739–767. (doi:10.1162/003355399556133)
45. Clauset A, Shalizi C, Newman M. 2009 Power-law distributions in empirical data. *SIAM Rev*. **51**, 661–703. (doi:10.1137/070710111)
46. Qomi MJ *et al.* 2014 Combinatorial molecular optimization of cement hydrates. *Nat. Commun*. **5**, 4960. (doi:10.1038/ncomms5960)
47. Qomi A, Javad M, Ulm F-J, Pellenq RJ-M. 2015 Physical origins of thermal properties of cement paste. *Phys. Rev. Appl.* **3.6**, 064010. (doi:10.1103/PhysRevApplied.3.064010)
48. Kilian L. 2007 *The economic effects of energy price shocks*. Social Science Research Network. See http://papers.ssrn.com/abstract=1140086.
49. Horowitz MJ, Haeri H. 1990 Economic efficiency v energy efficiency: do model conservation standards make good sense? *Energy Econ*. **12**, 122–131. (doi:10.1016/0140-9883(90)90046-I)
50. Pelenur MJ, Cruickshank HJ. 2012 Closing the energy efficiency gap: a study linking demographics with barriers to adopting energy efficiency measures in the home. *Energy* **47**, 348–357. (doi:10.1016/j.energy.2012.09.058)
51. Brand L. 1957 The Pi theorem of dimensional analysis. *Arch. Ration. Mech. Anal*. **1**, 35–45. (doi:10.1007/BF00297994)