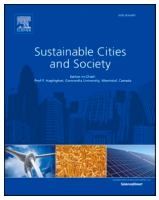
A bottom-up framework for analysing city-scale energy data using high dimension reduction techniques

Sustainable Cities and Society 89 (2023) 104323

Contents lists available at ScienceDirect

Sustainable Cities and Society

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



A bottom-up framework for analysing city-scale energy data using high dimension reduction techniques

Waqas Khan *, Shalika Walker, Wim Zeiler

Department of the Built Environment, Eindhoven University of Technology, Eindhoven, the Netherlands

ARTICLE INFO

Keywords:
Data mining
Dimensionality reduction
Machine learning
Feature importance
Data mapping

ABSTRACT

Worldwide cities are becoming more sustainable and are being monitored using data collection techniques at various geographical levels. Given the growing volume of data, there is a need to identify challenges associated with the processing, visualization, and analysis of the generated data from an urban scale. This study proposes a framework to investigate the capabilities of dimensionality reduction techniques (t-SNE, and UMAP) applied to city-scale data to identify key features of high consumption and generation areas based on building characteristics. The analysis is performed on measured data from 2735 postcodes consisting of 72000 households/buildings from a city in the Netherlands. The evaluation results showed that the UMAP's algorithm mean sigma quickly approaches a threshold of 0.6 at n_neighbor values of 50 and the low dimensional shape does not change with increasing values. Whereas the t-SNE's mean sigma value increases continuously with the increasing perplexity value, implying that t-SNE is significantly more sensitive to the perplexity parameter. The UMAP algorithm was used to extract information about the high photovoltaic generation and consumption regions. The proposed framework will assist grid operators and energy planners in extracting information from energy consumption data at the neighbourhood level by utilizing high dimensional reduction techniques.

1. Introduction

Countries are increasingly adopting long-term energy planning at the city level to reduce urban energy consumption and its associated emissions (Cajot et al., 2017). As the built environment is responsible for most of the primary energy use and carbon emissions. Therefore, understanding building energy use is a crucial component for advancing urban sustainability. Measures are needed to be implemented on multiple levels including energy-efficient appliances, efficient cooling and heating systems at the building level, and combined heat and power plants on the district-scale (Madlener & Sunak, 2011, Zhang et al., 2021). While the benefits are numerous, the goals are not easy to achieve as the distribution of energy demand and resources is rarely uniform at neighbourhoods and city levels (Damsgaard et al., 2015). There is a lack of understanding of the relationship between the distribution of energy systems and different demand characteristics from the building scale up to a city scale.

The built environment now has promising opportunities thanks to the development of smart metering infrastructure which measures and stores data on a building's electricity use at an hourly or sub-hourly resolution (Park et al., 2020). These electricity consumption datasets have the potential to support grid operators and policymakers to make informed decisions (Van Aubel & Poll, 2019). Therefore, countries are making an effort to collect energy consumption data and make it publicly available through open data platforms. In the Netherlands, more than 5.2 million smart meters were installed as of 2018 covering almost 54% of Dutch households (The Netherlands 2020). Big data from these households provide the potential for researchers, and utilities to better understand the energy profiles of buildings and utilize this information in cutting-edge applications (e.g., intelligent energy management systems, customer classification, portfolio analysis, and load profiling).

Big data analytics can be applied to the neighbourhood-scale energy consumption data to determine the optimal locations for windmills, biomass, and solar power plants as well as energy storage systems to negate the demand. It can also assist in understanding the impact of the adopted policies such as the promotion of solar panels on building rooftops. This data offers many opportunities to aid in the electrical transition between the built environment and the grid.

* Corresponding author at: Department of the Built Environment, Eindhoven University of Technology, De Zaale, PO Box 513, 5600 MB Eindhoven, the Netherlands.

E-mail address: w.khan@tue.nl (W. Khan).

https://doi.org/10.1016/j.scs.2022.104323
Received 24 June 2022; Received in revised form 23 October 2022; Accepted 22 November 2022
Available online 25 November 2022

Even though big data provide opportunities, these large datasets are complicated to analyse. Generally, studying the interactions of variables on a larger scale might become more time-consuming and ineffective in the smart building's energy field due to the increasing dimensions of data. The datasets can be highly disaggregated and can span multiple spatial levels from a single block to an entire neighbourhood or district. This significantly decreases the ability to extract information from the available energy data. Traditional tools and methods make it difficult to derive valuable insights from these continuously growing energy databases (Madlener & Sunak, 2011). High dimensional energy data sets are also often unstructured and noisy which makes it more difficult to apply appropriate data mining techniques. There is no general analytical approach that can provide insight into such problems. As algorithm tends to be problem and data specific. Therefore, there is a need to develop techniques to monitor and quantify changes in urban energy use patterns specifically from the perspective of high-dimensional data.

Understanding the energy use dynamics on neighbourhood scale is important for energy policy makers and city planners. It enables them to understand the energy use across the morphological contour of a city and provide contextual awareness for better allocating the resources and targeting policy measure to reduce overall consumption.

1.1. Review of the current literature

Previous studies have shown that data-driven machine learning and artificial intelligence techniques can make information extraction more efficient (Wang et al., 2014; Chen et al., 2018). They have adopted data mining techniques successfully for energy prediction (Sun et al., 2020; Khan et al., 2022), optimization and control (Barber & Krarti, 2022; Wang & Hong, 2020; Lee & Heo, 2022), classification (The Netherlands, 2020), occupancy detection (Rueda et al., 2020), thermal comfort (Ngarambe et al., 2020), fault detection and diagnosis (Zhao et al., 2020), and monitoring (Chen et al., 2018). These studies have focused on the evaluation and development of these models in an individual approach mainly on the building level and with a minor focus on the neighbourhood or city scale.

The available literature on neighbourhood scale analysis is comprised of two categories, namely, bottom-up and top-down approaches (Wei et al., 2018). The Bottom-up approach considers and aggregates individual building parameters at different modelling scales to extract relevant information. Whereas the top-down approach employs statistical data on the macro-level to determine energy behaviour. The top-down approaches represent the entire building block as a single unit for analysis purposes. While top-down approaches are easily implementable for a broad-scale analysis, bottom-up approaches can help identify valuable parameters for identifying focused neighbourhoods or regions in different spatial levels. The bottom-up approach has shown to be excellent for in-depth analysis based on current studies (Hong et al., 2020; Torabi Moghadam et al., 2017). Some studies have used data-driven models on a city scale to predict the energy use in buildings (Kontokosta & Tull, 2017), the impact on energy consumption due to residential density (Damsgaard et al., 2015), and discovering regions of different functions (e.g., educational areas, entertainment areas, and regions of historical interests) in a city (Wang & Hong, 2020)

From these neighbourhood-level bottom-up studies, only a few studies evaluated in detail the impact of other parameters on the energy demand of cities with real datasets. Most studies use linear approaches using simulated data to understand neighbourhood level data. The authors in (Brownsword et al., 2005) adopted a linear programming approach and utilized a wide typology to categorize urban energy "consumers," dividing the city into residential, commercial, and industrial applications in small, medium, and large size bins. While the authors' stated that the goal is for the model to be replicable, this method ignores the impact of these features on energy use. The authors in (Heiple & Sailor, 2008) attempted to simulate and scale-up building level estimation of the entire city of Houston, Texas using prototype buildings. The inclusion of prototypical buildings creates a significant error in the prediction, as the estimation adds up uncertainty in the prototype buildings and it is not comparable to real buildings. The simulated and actual energy consumption of building usually differ due to the characteristic and consumption pattern of buildings (Aranda et al., 2018). Building energy consumption is also dependent on user behaviour that can add more uncertainty to the real data and simulations are not able to take that into account. The predictive power of the citywide model is also limited due to a lack of data on actual energy use at the building level and building features.

Kontokosta (Kontokosta, 2015) investigated the causes of commercial building energy consumption across buildings, systems, geographic, and occupancy factors on actual building energy usage data for over 20, 000 structures in New York City. The author uses a multivariate regression model to discover that a building's annual total energy consumption and intensity are influenced by its size, age, usage, occupancy characteristics, construction type, and proximity to other buildings. The study indicates that relationships between parameters exhibit different patterns within different building classifications, and the assumption of linear correlations between variables often fails.

In (Zarco-Periñán et al., 2021) the authors segmented cities based on density into five groups. They investigated the thermal and electrical consumption for each group based on the density of inhabitants. The findings obtained on consumption per household showed that the higher consumption is in the densely populated cities, except for four cities, in which consumption decreases slightly compared to the other groups. Furthermore, in the case of thermal consumption, they got varied results for different groups of cities. Their approach only provides the base information and is unable to link additional parameters to discover any connections between the underlying reasons for the variations.

1.2. Problem identification

The majority of previous work relies on techniques that uses simulated data for neighbourhood level energy data analysis approaches (Hong et al., 2000; Ferrando et al., 2020). There is also a lack of focus on evaluating the consistency of algorithms and generalizing their results for extracting information from high dimensional data. Moreover, many of these data-driven models used for analysing neighbourhood and city scale data evaluated linear regression models or individual visual inspection of the parameters (Aranda et al., 2018). With the increasing number of variables, the uncertainty and complexity of the data are increasing. It is difficult to understand and interpret the fundamental behaviour of the condition being observed through user exploration of big-time series data. There is a need to minimize the dimensionality of the data by reducing the number of variables or features and focusing on a lower-dimensional subspace that captures the essence of the data to provide adequate information and insights at different spatial levels.

Non-linear dimensionality reduction techniques are used in other fields to avoid the problem of overcrowding and to extract valuable information from high-dimensional data (ABC, 2022). Specifically, the t-Distributed Stochastic Neighbour Embedding (t-SNE) (van der Maaten & Hinton, 2008) and Uniform manifold approximation and projection (UMAP) (McInnes et al., 2020) algorithms are the most commonly used techniques in the literature for different applications (ABC, 2022). Both these techniques have certain limitations and advantages. The relative behaviour of the techniques is different in other fields and needs to be evaluated in detail for the building energy data (van der Maaten & Hinton, 2008).

To the best of the authors knowledge, high dimensional data on building energy data is not analysed on a neighbourhood scale due to its size and its nonlinear nature. There is also lack of focus on a complete framework development from the data pre-processing to post mining of the data. Moreover, dimensionality reduction techniques are barely used in the literature for the visual inspection of neighbourhood scale energy data mining specifically linking building use characteristics with the

energy data. This work aims to fill the identified research gaps with the following objectives:

- Explore the emerging need for high dimensional reduction techniques for visual analysis of building characteristics and energy data
- Evaluate the key high-dimensional reduction methods to analyse building energy data at a neighbourhood or city scale and the insights that this unsupervised workflow can provide
- Cross-compare the applications, strengths, and limitations of the selected methods and find the best way to characterise low-dimensional behaviour
- Provide confirmatory procedure for the obtained results

These objectives are achieved by developing a framework to evaluate high-dimensional analysis techniques. The used framework can indicate the relationship between variables (for example building energy consumption and building characteristics). This work will provide a methodological contribution to the existing literature on urban energy analysis and planning from three perspectives. First, this work aims to present a framework that can establish the relationship between the urban energy consumption profile with the building characteristics to investigate possible trade-offs and interactions that have not yet been sufficiently explored. Second, it explores non-linear dimensionality reduction techniques from the perspective of local and global preservation that is seldomly used in the literature. Third the framework will provide logic behind the mapping of dimensionality reduction to confirm the validity of the obtained results.

The framework will have the potential for energy system to analyse data in a neighbourhood-based approach in a spatial context. The research will provide the basis for exploring urban energy data on a neighbourhood level to support informed interventions, suggest strategies, and identify priorities and the most suitable location for zero-energy neighbourhoods and district developments.

The rest of the paper is organized as follows. Section 2 introduces the proposed framework and the sub-steps required which includes the introduction of the case study, the proposed methods, and evaluation techniques. Sections 3 and 4 provide the evaluation results of the framework and discuss the outcomes in detail. Section 5 focuses on summarizing the overall outcome of this research.

2. Description of the proposed framework

The flow diagram of the proposed framework is shown in Fig. 1. The flow diagram is divided into four steps based on different outcomes of this work and a more detailed description of each step is provided.

- Step 1 provides the details about the data collection and pre-processing phase. Here, the data is cleaned and manipulated for the functioning of the next steps.
- Step 2 focuses on the visual analysis of the original data and the evaluation of input features using feature correlation and normality test
- Step 3 evaluates and compares high-dimensional reduction techniques based on the ability of global shape preservation and reusability.
- Step 4 extracts high consumption and generation regions of interest and uses feature association to underline the relevance of the extracted features.

2.1. Data acquisition and pre-processing

This section describes the data collection and pre-processing steps used in this study. Fig. 2 presents the flow diagram of this phase. The neighbourhood scale energy data collection process for a whole city is time intrinsic and difficult due to privacy issues. However, it was made possible due to different open-source public datasets and the project partner. A city in the Netherlands is used as the case study in this research with the year 2020 as the reference period. The city wants to be energy neutral by 2044 and explore new ways to reduce its energy consumption and implement sustainable energy sources. As a result, the city has started taking initiatives, such as giving incentives for the installation of solar panels and installing large-scale photovoltaics and wind turbines in business parks. The city officials have also joined forces with individuals, businesses, and research institutes to highlight focused areas to encourage new efforts to further reduce carbon emissions. In this context, it was of high interest for them to explore new ways of extracting information from the neighbourhood scale energy data.

2.1.1. Data collection

This paper uses a bottom-up approach and combines datasets from several sources as neighbourhood scale energy datasets are not publicly available. Smart meters and services data for households are obtained from two different sources. The smart meter data is obtained from the energy provider in the city. The spatial dimension is based on the aggregation of household data per postcode. The data is aggregated for a

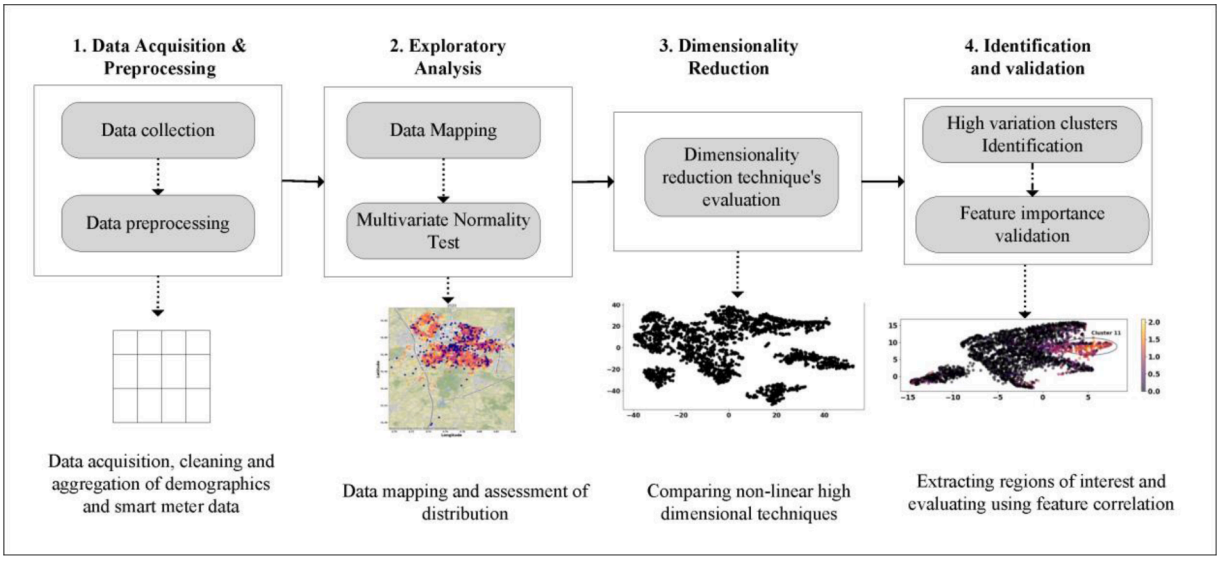


Fig. 1. A flow diagram of the proposed framework.

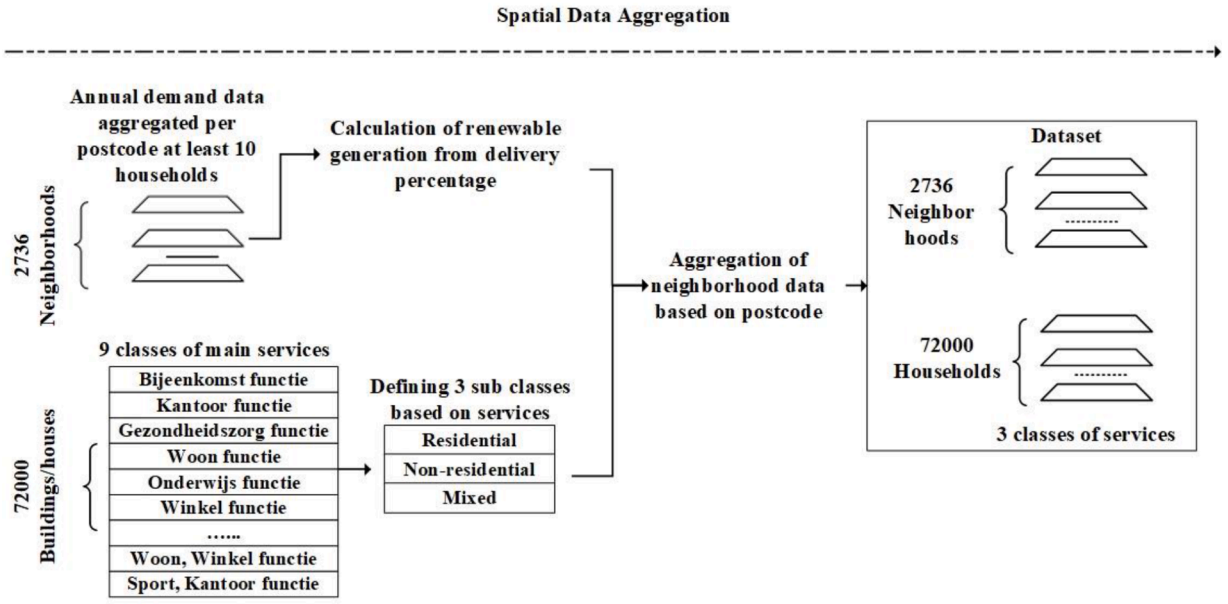


Fig. 2. Flow diagram of the data gathering and pre-processing phase. The gathered data contain two datasets: (1) smart meter aggregated data of households based on postcodes(neighbourhoods) (2) individual household registration data of the number of households and services in 2736 neighbourhoods. Two data processing steps are described. First, for the smart meter data, self-production is calculated for each of the neighbourhoods from delivery percentage and annual consumption. Second, the service-based household data is classified into three main classes based on functionality. A single dataset of mutual neighbourhood level is created.

minimum of 10 connections based on the postcode. If a postcode has a smaller number of connections, they are aggregated with another postcode's data.

Building service data are based on the Centraal Bureau voor de Statistiek (CBS) Netherlands (voor de Statistiek, 2022). This is a publicly available resource that provides detailed service-based data per household's functionality. The individual smart meter data and neighbourhood level are protected by privacy laws; therefore, the scale of the data is described using a categorical threshold using low, medium, high, and very high based on the capacity of the variable in the visualisation phase.

2.1.2. Data pre-processing

The main data pre-processing step in this phase is cleaning, which includes removing missing data and duplicating neighbourhood and household data depending on the postcode for both datasets. Following data cleaning, both datasets are made compatible to be easily merged. However, before making the individual household's service and aggregated smart meter data compatible the following pre-processing is performed on the individual datasets:

- For each postcode, the self-production is collected in delivery percentages, where 100% means that no energy was given back to the grid. Actual self-production is calculated from delivery percentage and annual consumption.
- The building characteristic dataset registered the households into 9 main classes based on their functional purpose. However, some households or buildings are registered as mixed entities. For instance, a single building can accommodate multiple functional entities e.g., a single office building can be used by multiple companies, that are all registered at the same address. For analysis purposes, the buildings are classified into three main subclasses: "Residential", "Non-residential", and "Mixed" building types.

The aggregated smart meter data is then combined with building characteristic data based on postcode to make the datasets compatible for use in subsequent phases of the research. The result is a combination of two mutually compatible datasets that describe the aggregated spatial energy smart meter data of 72000 households and services in 2736 neighbourhoods. The neighbourhoods refer to each of the postcodes. The combined spatial dimension dataset consists of 8 variables describing different energy and demographic attributes of a neighbourhood namely "Number of connections (Number of houses/buildings connected to the grid)", "Consumption", "Self-production (PV generation given back to the grid)", "Smart meter percentage", "Annual consume low tariff (Consumption during non-peak hours)", "Residential", "Non-residential", and "Mixed". Each of the parameters is important in studying different applications of the energy transition. For instance, the "Number of connections" parameter is important in specifying the number of households/buildings connected to the grid in a neighbourhood. The "Annual consume low tariff" can assist in investigating whether people are using more energy during the off-peak hours to reduce their energy bills.

2.2. Exploratory analysis

2.2.1. Data mapping

Following the data collection and pre-processing, data mapping is used to visualise the different high-dimensional parameters of the city individually. The data visualisation is based on the longitude and latitude data that is linked with the postcode data. The amount of data points is very large based on the neighbourhoods which can result in an overlapping cluttered representation that does not effectively inform about the density and the structure of points. Hexagonal binning is used for visualizing data with a more implied structured gesture (Lewin-Koh, 2021). Instead of rendering a scatterplot of thousands of points, hexbinning the points into a few hundred hexagons can imply general distribution. The city map visualised based on a hexagon tessellation constructed over a specific region for the overlapping neighbourhoods can result in a clear visualisation. The number of neighbourhood data points falling in each hexagon is counted and stored in a data structure. The hexagons are represented by the average colour from all the data points using a colour gradient. As colours are represented numerically in computers. The intensity of the variables falling in a specific hexagon can be represented by the average numerical value. The use of a hexagon tessellation at such a spatial resolution has a long history in spatial

analysis and is supported by similar urban analyses (Burdziej, 2019). An example of the overlapping hexagon scatter plot by hexbinning can be observed in Fig. 3.

To link the neighbourhood data with the geographical distribution of the city, a base map is added to the hexagonal representation (ABC, 2022). The term "base map" refers to a collection of geographic data that serves as the mapping background. A base map serves as a background for other layers that are superimposed on top of it. Base maps are used to locate city characteristics that do not change frequently, such as roads, highways, rivers, and boundaries. This information is usually contained in a base map, and then extra layers with specific data from a specific discipline are overlaid on top of the base map layers for analysis purposes. The base map is added to the visual inspection of the aggregated energy and building characteristics data for comparison purposes.

2.2.2. Data validation test

The data validation test plays an important role in the selection of high-dimensional data reduction techniques. Several models are built on the assumption that the data variables have a normal distribution and are linearly related. For this reason, the linearity between variables is evaluated using the Pearson correlation method (Berman, 2018) which measures the strength of linear association between variables.

Following the correlation analysis, the Anderson-Darling Test is employed to evaluate whether a data sample comes from a normal distribution due to its capability of returning a list of critical values rather than a single p-value (Nelson, 1998). This test is used to assess the distribution of a dataset acknowledged by many researchers (Jäntschi & Bolboacă, 2018). This test provides a basis for a more thorough interpretation of the obtained results.

The critical values in the test are at a range of pre-defined significance boundaries at which the formulated hypothesis can be rejected. The critical values depend on the distribution the data is tested against. This test works for normal, logistic, exponential or Gumbel distributions. The critical values for a normal distribution are based on the significance level of 1, 2, 5, 10 and 15 per cent (ABC, 2022).

The two hypotheses formulated for the normal distribution test are:

- H0: The data follows the normal distribution
- H1: The data do not follow the normal distribution

The hypothesis that the distribution is of a specific form is rejected if the test statistic value is greater than the critical value.

Based on the results of the Pearson correlation and Anderson Darling test, the selection of linear or nonlinear high dimensional reduction techniques can be made. If data variables have a linear relationship and follow a normal distribution, linear dimensionality reduction techniques can be used, and vice versa.

2.3. Dimensionality reduction

Dimensionality reduction (DR) is one of the main techniques for reducing redundant features, and analysing high-dimensional data (Vachharajani & Pandya, 2022) to improve the model's feature learning accuracy. This is achieved by adjusting an objective function by exploiting redundancy between variables and producing a reduced new set of variables. It can be defined as a set of techniques that map high-dimensional datasets to a low-dimensional subspace while retaining as much original information as possible from the original data. They can also help in identifying the most relevant feature for further research or data representation. Dimension reduction techniques are mainly divided into two major categories: linear dimension reduction and nonlinear dimension reduction (manifold learning).

- The linear reduction techniques use linear transformation to project the data from high-dimensional to lower-dimensional space. Principal component analysis (PCA) is one of the most popular linear dimensional reduction algorithms used in the literature (El Bouchefry & de Souza, 2020). PCA reduce the dimensionality of data that is highly correlated by transforming the original data to a new set which is known as the principal components (eigenvectors). PCA extracts the low dimensional axes by reorienting the high dimensional data by maximizing the variance of the variable.
- The non-linear dimensional reduction techniques use manifold learning to project the data from high dimensions to lower-dimensional space. The most popular non-linear dimensionality reduction techniques are t-SNE and UMAP (Jiale & Ying, 2020).

In recent years, machine learning algorithms such as the t-SNE algorithm and the UMAP algorithm have become especially popular (Demidova & Stepanov, 2020; Pal & Sharma, 2020). The focus of this research is mostly on non-linear dimensional reduction techniques for building energy data due to its complexity and nonlinear nature described in more detail in Section 3.2.

2.3.1. t-SNE

t-SNE is a non-linear dimensionality reduction technique proposed by van der Maaten and Hinton for high-dimensional scaling in 2008 (van der Maaten & Hinton, 2008). It is widely used in genome data (Kobak & Berens, 2019), hyperspectral imaging analysis (Pouyet et al., 2018), and word processing (van der Maaten & Hinton, 2008) for the reduction and visualisation of high-dimensional datasets. This technique is popular due to its exceptional ability to scale high-dimensional data to

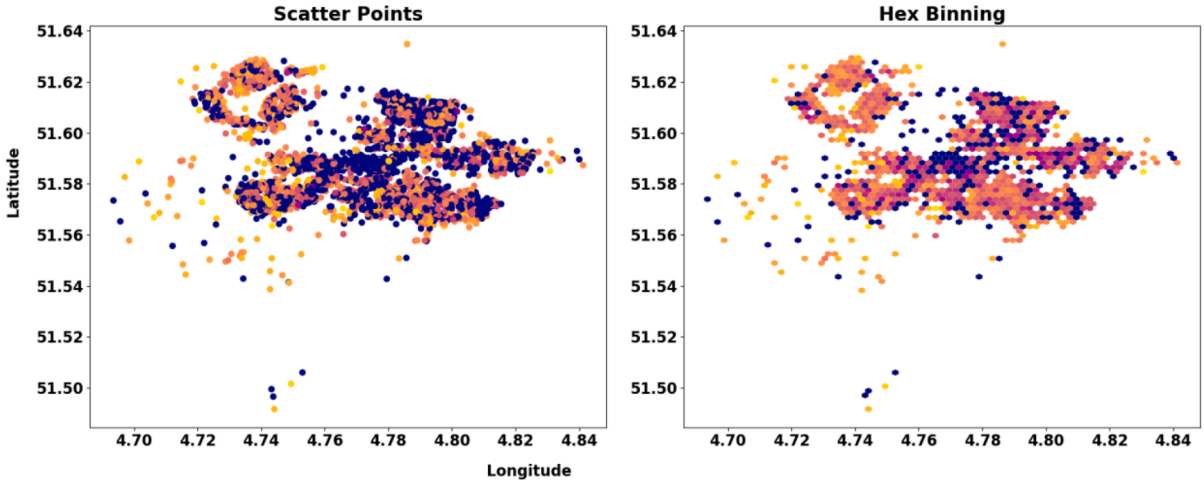


Fig. 3. Illustration of scatter points on top of each other (left) and hexbinning representation (right).

lower dimensions.

The algorithm calculates the Euclidian distances of each point from all the other points. Then, take these distances and start determining the conditional probability of similarity of points in high-dimensional space. Let's consider a data set that contains n samples $x_1, \ldots, x_n$. The goal is to find a low dimensional mapping points $y_1, \ldots, y_n$. The conditional probability of a point x_j given x_i can be mathematically calculated in the following way (van der Maaten and Hinton, 2008):

$$p_{j|i} = \frac{exp^{\left(-\|X_i - X_j\|^2 / 2\sigma_i^2\right)}}{\sum_{k \neq i} exp^{\left(-\|X_i - X_k\|^2 / 2\sigma_i^2\right)}} \quad (1)$$

Where x_k is the random projection of point x_i. The probability of point x_j to be next to point x_i is represented by a Gaussian cantered at x_i with a standard deviation of σ_i. The σ_i is controlled with the predefined parameter; perplexity (van der Maaten & Hinton, 2008). The σ_i is important in finding the optimum for the model to converge and vary with the perplexity parameter.

These conditional probabilities are symmetrized to obtain joint probabilities defined as (van der Maaten & Hinton, 2008)

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n} \quad (2)$$

Where n represents the total size of the dataset.

In the next step, the t-SNE employs the student t-distribution (van der Maaten & Hinton, 2008) with a single degree of freedom to avoid overcrowding. A t-distribution curve is like a normal distribution; however, it is shorter and has a fatter tail. With this distribution, it calculates the probability of similarity of points in the corresponding low-dimensional space. The t-SNE calculates the similarity probability distribution q_{ij} as (van der Maaten & Hinton, 2008)

$$q_{ij} = \frac{\frac{1}{1+\|y_i - y_j\|^2}}{\sum_{k \neq l} \frac{1}{1+\|y_k - y_l\|^2}} \quad (3)$$

It then tries to minimize the difference between these conditional probabilities (p_{ij} *and* q_{ij}) in higher and lower-dimensional space for a perfect representation of data points in lower-dimensional space. The t-SNE measure the similarity between two probability distributions using the Kullback-Leibler divergence cost function (KL) (Kingman, 1970) and is given by (van der Maaten & Hinton, 2008)

$$KL\ (P|Q) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (4)$$

Where P and Q are the distributions for p_{ij} and q_{ij} in the high and low dimensional space respectively.

After calculating the two probability distributions describing the high and low dimensional space. The cost function is optimized using gradient descent. The t-SNE method uses the Barnes-hut approximation method as a gradient calculation algorithm (ABC, 2022; van der Maaten, 2014) and it approximates the gradient at each iteration to adapt to the data. The Barnes-hut method is significantly more scalable. It can embed thousands of data points before becoming computationally intensive.

The t-SNE also performs a binary search for the values of σ_i resulting in a value of p_i with user-specified perplexity value which is defined as (van der Maaten & Hinton, 2008)

$$Perplexity(p_i) = 2^{H(p_i)} \quad (5)$$

Where H (p_i) is Shannon's entropy of p_i measured in bits.

$$H(p_i) = \sum_j p_{j|i} \log_2 p_{j|i} \quad (6)$$

In this way, t-SNE maps the high-dimensional data to a lower-dimensional space and attempts to extract information from the data by extracting observed clusters based on similarity from data with multiple features. However, after this process, the data is no longer identifiable, and any inference based only on the output of t-SNE cannot be made. Therefore, it is mainly a data reduction and exploration technique. For the detailed working of the t-SNE algorithm, the readers are referred to (van der Maaten & Hinton, 2008).

2.3.2. UMAP

UMAP is a nonlinear dimensionality reduction technique (McInnes et al., 2020) similar to t-SNE that can be used to reduce high dimensional data. UMAP generates a low-dimensional graph of data that preserves the cluster representations of the high-dimensional data and their relationship to each other (McInnes et al., 2020). Unlike t-SNE, UMAP can better preserve the global structure of the input data in the low-dimensional space. In the t-SNE representations of low dimensional data, the within-cluster distances are meaningful for determining the similarity of data, but the distances between clusters are not guaranteed to be significant. This improvement is critical because real-time data clustering relies on the separation of data in the low-dimensional space to find meaningful relationships.

UMAP can be divided into three major steps as illustrated in Fig. 4:

a) Calculation of distance of high dimensional points
b) Construction of a high-dimensional graph from the data
c) Mapping to a low-dimensional representation

The first step in UMAP is the calculation of distances between each pair of high-dimensional points and the second step focus on plotting the distances on a graph. The points that are far away from the initial point will be plotted further away on the graph. This is done by drawing a curve over the data to calculate the similarity scores. The high dimensional data points similarity score (HSS) is calculated using the following equation:

$$\sum_{j=1}^{k} \exp\left(\frac{-max\left(0, d\left(x_i, x_{ij}\right) - p_i\right)}{\sigma_i}\right) = \log_2 k \quad (7)$$

Where p_i is the distance to the first nearest neighbour from the point x_i raw data, $d(x_i, x_{ij})$ is the distance between x_i and x_j and σ specifies the shape of the curve drawn, similar to the perplexity in t-SNE. The σ depends on n-neighbour's parameter that is specified beforehand. UMAP scales the curves by changing the sigma value so that regardless of how close or far the neighbouring points are, the sum of the similarity scores will be equal to the specified log2(number of neighbours). After the similarity score calculation, UMAP initializes a low-dimensional graph. UMAP uses spectral embedding to initialize the low dimension graph to place the points closer to each other in the same low dimensional clusters as the high dimensional visualisation. UMAP randomly select a pair of points based on probabilities in a cluster, proportional to their high-dimensional score. UMAP again calculates low-dimensional similarity scores however instead of using a variety of curves like the high-dimensional data, the low-dimensional similarity scores come from a fixed bell-shaped curve that is derived from a t-distribution (van der Maaten & Hinton, 2008). UMAP uses the following equation to calculate the low- dimensional similarity scores (LSS) (McInnes et al., 2020; Jiale & Ying, 2020):

$$LSS = 1 + \propto d^{2\beta} \quad (7)$$

Where d is the distance between two low dimensional points and α and β are by default set to 1.93 and 0.79 respectively (Jiale & Ying, 2020). The α and β values can be modified with user-defined parameters for the minimum distance between low dimensional points and their spread. It gives more control over how tightly low-dimensional points end up. UMAP uses Stochastic gradient descent to find the optimal low-dimensional graph. Here the basic details of the functioning of

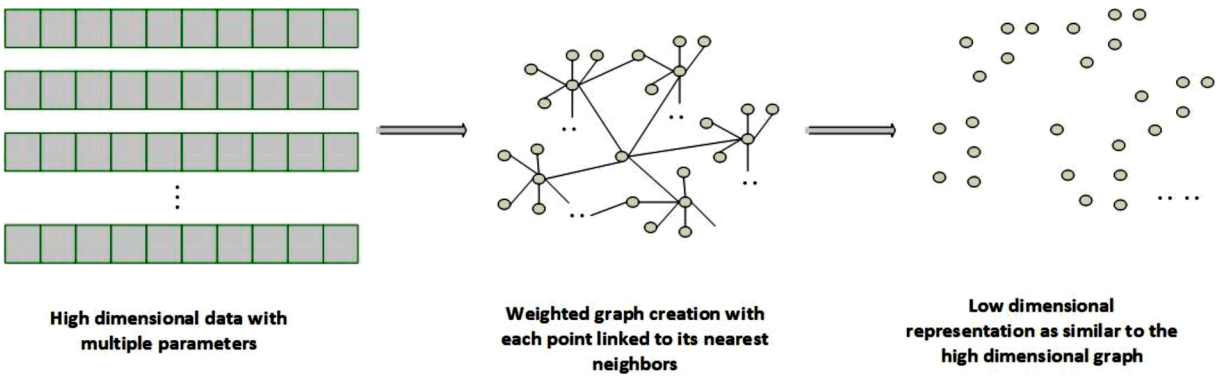


Fig. 4. UMAP algorithm working principle graphical representation

UMAP are provided however for a complete description of the UMAP algorithm, please see (Jiale and Ying, 2020).

2.3.3. Comparison between t-SNE And UMAP

In this study, the non-linear dimensionality reduction methods t-SNE and UMAP have been used. Although both these algorithms are useful in their way, there are certain advantages and limitations of these algorithms. The t-SNE algorithm has an issue in mapping data that leads to potentially misleading results (ABC, 2022). The cost function in Eq. (4). used by the t-SNE algorithm only preserves the local structure of the data, while not maintaining the global structure which can lead to false intuitions (ABC, 2022). On the other hand, UMAP can preserve more of the global structure compared to the t-SNE. The main reason why the t-SNE does not preserve global structure is also linked with its random initialisation. UMAP uses a Laplacian Eigenmap for initialisation (McInnes et al., 2020) that leads to the preservation of a more global structure.

It is also possible in t-SNE algorithm to increase the perplexity value to an extent that the algorithm considers more neighbouring points that could lead to the preservation of the global structure. As a result, t-SNE and UMAP may act similarly in this regard. However, at a very large perplexity value t-SNE start behaving like a multi-dimensional scaling algorithm (ABC, 2022). Therefore, the t-SNE algorithm shows only linear global structures that are not well suited for non-linear real-world data.

On the other hand, UMAP uses cross-entropy in the cost function that enables large penalties if small distances between data points are mapped to large distances in the low dimensional representation or vice versa. This guarantees that data points that were similar in the original high-dimensional space would stay similar in the low-dimensional embedding (see Fig. 4). As a result, in the low-dimensional embedding, the UMAP technique will maintain more of the high-dimensional global structure.

Another disadvantage of the random initialization of the t-SNE algorithm is its lack of reusability. It means that running t-SNE on the same dataset twice will result in different low dimensional data initialization. This generally decreases the reproducibility of the obtained results.

Despite all the arguments, the main reason for the different ways of t-SNE and UMAP global structure preservation is the approach employed to solve the minimization of the cost function. While t-SNE uses gradient descent, UMAP employs stochastic gradient descent. The gradient is computed only for a subset of the total dataset in the latter. This improves overall processing speed while lowering memory use. However, adopting this method has the disadvantage of a lower convergence rate than using normal gradient descent. Overall, UMAP outperforms t-SNE in terms of speed, mathematical background, global structure preservation, and memory consumption.

2.4. Identification and validation

This section aims to evaluate and validate the results of high-dimensional visualisation algorithms. The first part focus on identifying the density of regions with high consumption and self-production from the perspective of residential, non-residential, and mixed buildings. The second part focuses on the validation of the obtained results by extracting the feature correlation from the obtained high-dimensional representation.

2.4.1. High variation regions identification

Extraction of regions with similar patterns is an essential task of high dimensional reduction techniques. Similarity identification helps in understanding the overall trend in a data. If dimensionality reduction can find similar patterns in a dataset, we can identify the underlying relationship for that specific area. Clustering is used in the literature on the obtained target variable from the high dimensional reduction techniques to group similar points together (Yang et al., 2021). Clustering classifies data into groups to discover latent features within each group. A good cluster visualization allows the user to visualize groups of data points easily. There should be enough space between the groups in scatter plots such that points in the same group are closer to each other than those in other groups.

However, the problem with high dimensional reduction techniques is that they do not preserve density or distance (Nguyen & Holmes, 2019). They only preserve the nearest neighbour's points structure to some extent. The distinction is very subtle, but it has an impact on the result of distance- or a density-based algorithm like k-means. While clustering can sometimes work, it's impossible to know whether the discovered "clusters" are actual or just dimensionality reduction. To tackle this problem, the obtained target variables from the high dimensional reduction technique are mapped (coloured) with the original data variables in this research.

This visualisation will help in understanding the relations of the input data with the clustered points in the low dimensional space. Mapping the low-dimensional data with the target variables will result in an accurate estimation of the high-dimensional features. The regions are then extracted based on the intensity of the important parameters (high self-production and consumption areas) for further evaluation.

2.4.2. Feature importance

Following the identification of the potential regions, the proposed workflow proceeds to use feature importance. It is a highly relevant task in data-driven information discovery techniques. Feature importance is used to describe what each region represents in the context of the input data. As high dimensional visualisation techniques cluster similar points in a similar location, the association of the input parameters is important in placing points closer to each other. Several parametric and non-parametric measures are used in the literature for association such as

Pearson correlation, linear regression and mutual information score etc (Song et al., 2012).

In this research, the Maximal information coefficient (MIC) is employed due to its ability to measure the linear and non-linear association between the input variables and the obtained low-dimensional target variables (Reshef et al., 2011). This method is used due to its superiority in capturing a wide range of relationships between variables (Reshef et al., 2011). It is based on concepts from mutual information theory. It ranges from 0 to 1, with 0 indicating independence and 1 indicating a noiseless functional association. The purpose of MIC is equitability: regardless of the type of relationship, similar scores will be seen in relationships with equal noise levels. As a result, finding a smaller group of the strongest connections may be particularly valuable in high-dimensional contexts. Whereas distance correlation may be better at detecting the presence of dependencies, the MIC is more geared toward assessing the strength and detecting patterns that might otherwise be missed by visual inspection (Reshef et al., 2011).

Feature importance can help in identifying the influencing parameters on the local demand characteristics from a large subset for further evaluation from a large number of variables. It will make the high-dimensional problem more concise and help other in comprehending the problem better.

2.5. Software environment

All the models in this study are developed using python version 3.8.12 in the Spyder development environment. Data mapping is done using matplotlib version 3.5.0. Contextily package is used to retrieve the tile base maps. The base map is used for data mapping with the matplotlib figures (ABC, 2022).

The t-SNE algorithm is implemented in Scikit-learn version 1.0.2 (ABC, 2022). The UMAP algorithm is implemented from the original GitHub repository by McInnes (McInnes, 2022). The mean sigma for the t-SNE and UMAP algorithm is evaluated from the work of Oskolkov (Oskolkov, 2022) on the energy dataset. The maximal information score for the feature importance evaluation is implemented from the original GitHub repository by Albanese et al (ABC 2022; Albanese et al., 2013).

3. Results

The following subsection focus on the implementation of the proposed framework in Section 2. Each of the subsections explains the findings of the applied framework in detail.

3.1. Data visualisation using mapping

In the first phase of the project, data mapping is used to visualise all the variables individually presented in Fig. 5. This visualisation is performed to provide an initial assessment of the high-dimensional data. The data intensity is anonymized due to privacy issues and is divided into four categories from low to very high. The data points are plotted on top of a base map of the city. The x and y-axis represent the longitude and latitude information based on the postcodes. The data points are aggregated and visualised as hex bins due to the large amounts of overlapping data points. Based on a preliminary analysis of the variable it is difficult to find any relation between the consumption and self-production with the other variables (number of connections, low-tariff consumption, smart meters percentage, residential, non-residential and mixed). There may be no visible relationship between variables, but they may be dependent. Also, as the amount of data point is very high visual analysis does not return any meaningful results.

Some general observations can be made that the consumption is higher in the city centre, whereas the self-production is higher more on the outskirts of the city. The non-residential building characteristics do not show any association with the high consumption or self-production areas. The smart meter percentage and annual low tariff consumption are almost identical in all the neighbourhoods except the city centre. Both these features could also be highly correlated to each other and may not be of any additional information for the objective of this study which is to evaluate the building characteristic effects on overall consumption and self-production.

3.2. Data validation test

After the initial data exploration, the data variables are evaluated to select the appropriate class of high dimensional techniques. The high dimensional data is evaluated using the correlation analysis and Anderson-Darling Test. Table 1 presents the correlation coefficient values of all the variables with each other. There is almost no linear correlation of the parameters with the consumption and self-production except for the non-residential buildings which show a slight correlation with the consumption data. Overall, the correlation analysis shows that there is no linear relationship between most of the parameters. However, this simple statistical analysis does not provide any significant

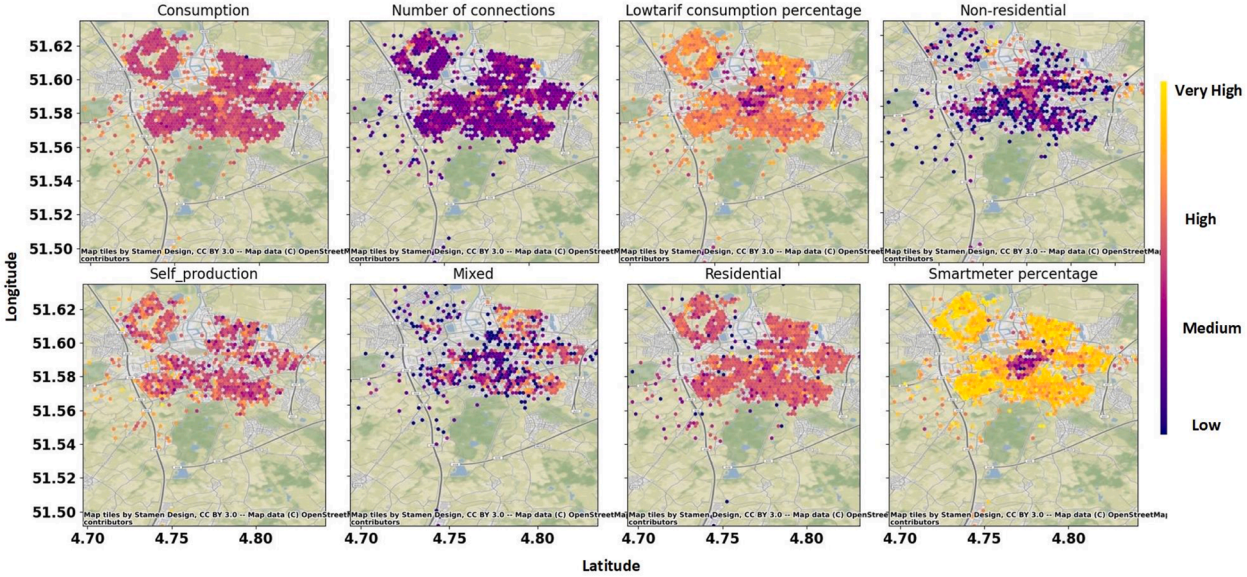


Fig. 5. Original data variables mapping using Hex Binning and adding base map for initial visual analysis.

Table 1
Correlation analysis of the original feature in the high dimensional space.

Parameters	Number of connections	Consumption	Self production	Mixed	Non-residential	Residential	Smart meter percentage
Number of connections	1						
Consumption	-0.11	1					
Self production	-0.087	0.076	1				
Mixed	0.007	-0.040	-0.054	1			
Non-residential	0.047	0.467	0.044	-0.002	1		
Residential	0.658	-0.164	-0.116	0.020	-0.025	1	
Smart meter percentage	-0.016	-0.222	0.101	0.033	-0.203	0.004	1
Low tariff percentage	0.013	-0.283	0.094	0.038	-0.319	0.061	0.738

detail. This coefficient is not a perfect indicator of data distribution; it just detects the linear correlation. However, it is an easy way to calculate the dependency among variables.

To investigate the distribution of the data, the Anderson Darling test is applied in this research. Table 2 provides the critical values at the pre-defined significance boundaries. The critical values are similar for all the variables because of the same sample size. To check if the test findings are significant, the statistic value is compared to the critical value. If the test statistic value is greater than the critical value, the results are significant, and the data distribution is not normal. The test results are significant for all variables, and the data distribution for the variables is not normal regardless of the degree of significance chosen. The H0 hypothesis (normal data distribution) can be rejected.

The correlation and the data distribution test show that there is almost no correlation between the parameters and the data distribution is not normal implying that using linear high dimension reduction techniques will not generate clear results. Techniques like PCA reduce the dimensionality of data that is highly correlated by transforming the original data into target variables to retain as much information as possible. Due to the non-linearity and non-normal distribution of the energy dataset this research mostly focused on evaluating non-linear reduction techniques to represent the data in low dimensional space.

3.3. High dimensional reduction algorithms comparison

This section compares the results of the t-SNE and UMAP algorithm in the context of global structure preservation and reusability. Both algorithms are used to map high-dimensional input data to low-dimensional 2D space. The algorithms are evaluated for different values of the hyperparameters of t-SNE and UMAP: perplexity and n-neighbours respectively. These hyperparameters define the balance between local and global aspects of the data. Fig. 6 and Fig. 7 show the obtained results for both algorithms. In the reduced representation, the coordinate axis of UMAP and t-SNE algorithms have no significant relevance, however, they are included in the representation to identify high consumption and self-production regions in the next phase.

From Fig. 6 it can be observed that the low dimensional representation by the t-SNE algorithm is not consistent, and the point representation change with each value of the perplexity parameter.

On the other hand, the low dimensional representation by UMAP in Fig. 7 shows consistent results after the n-neighbours value of 50. After that the increase in "n_neighbours" value is making the data point's getting closer to each other without affecting the overall shape of the low dimensional space.

In both techniques, the mean sigma (σ) parameter is mainly responsible for deciding how much data points can feel each other. The means sigma value is based on the hyperparameters of t-SNE and UMAP: the perplexity and neighbours respectively that decide the presence of its closest neighbours. It is possible to evaluate how the mean sigma value changes with different hyperparameter values for both algorithms. Fig. 8 shows the mean sigma dependency on the perplexity and n_neighbours. By increasing the n_neighbours hyperparameter, the UMAP's algorithm mean sigma quickly approaches a specific threshold and finds an optimum value. After the specific threshold, it is unaffected by the increasing n_neighbor values. This results in a similar representation in the low dimensional space of the original variables in every iteration.

Whereas the t-SNE's mean sigma value increases slowly and continues to increase with the increasing perplexity value, implying that t-SNE is significantly more sensitive to perplexity. The increasing mean sigma value of t-SNE at large perplexities has a profound impact on the gradient of the cost function that leads to the different formulations in the low dimensional space.

Another main difference between the two techniques was the computational efficiency. The t-SNE algorithm was very slow compared to the UMAP with similar hyperparameter values and the same amount of data.

Based on the global optimum reachability and the computational time capability, it is evident that the UMAP algorithm has consistent results and is more suitable for extracting information from high-

Table 2
Anderson Darling statistics for evaluating the distribution of parameters for the selection of the high dimensional technique.

Variables	Critical values at pre-defined significance boundaries				Statistic		Data distribution
Consumption	1	2	5	10	15	417.6	data does not follow a normal distribution
	1.09	0.91	0.78	0.65	0.57		
Number of connections	1	2	5	10	15	282.2	data does not follow a normal distribution
	1.09	0.91	0.78	0.65	0.57		
Low tariff percentage	1	2	5	10	15	106.811	data does not follow a normal distribution
	1.09	0.91	0.78	0.65	0.57		
Self-production	1	2	5	10	15	367.8	data does not follow a normal distribution
	1.09	0.91	0.78	0.65	0.57		
Mixed	1	2	5	10	15	566.0	data does not follow a normal distribution
	1.09	0.91	0.78	0.65	0.57		
Non-residential	1	2	5	10	15	600.8	data does not follow a normal distribution
	1.09	0.91	0.78	0.65	0.57		
Residential	1	2	5	10	15	182.1	data does not follow a normal distribution
	1.09	0.91	0.78	0.65	0.57		
Smart meter percentage	1	2	5	10	15	192.2	data does not follow a normal distribution
	1.09	0.91	0.78	0.65	0.57		

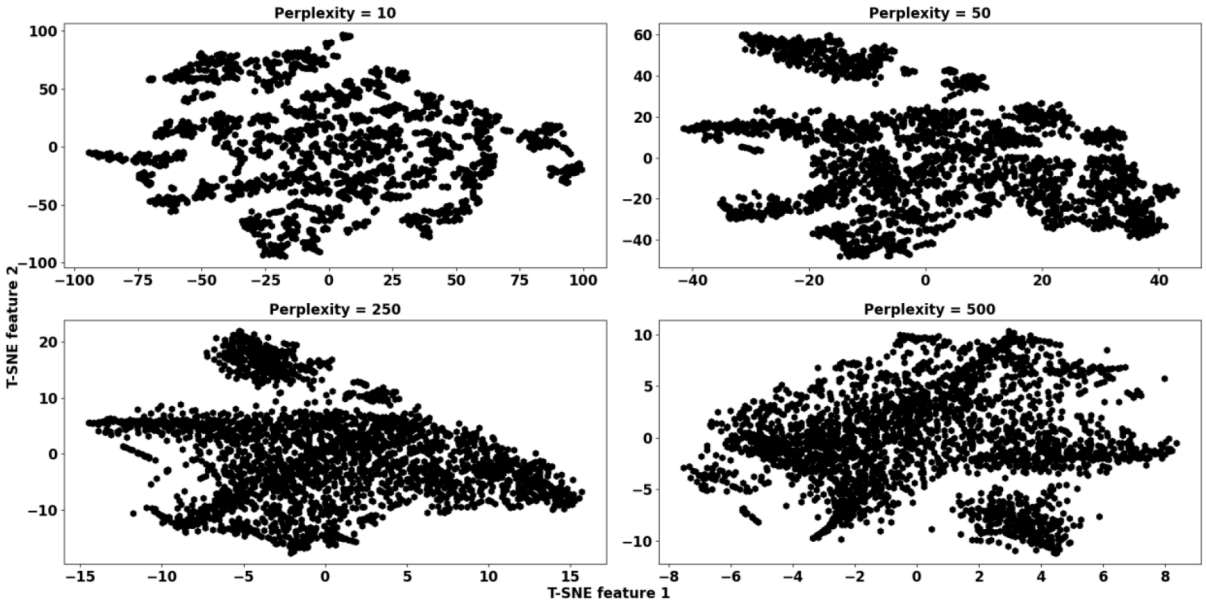


Fig. 6. Low dimensional representation of the original data by the t-SNE algorithm with different perplexity values

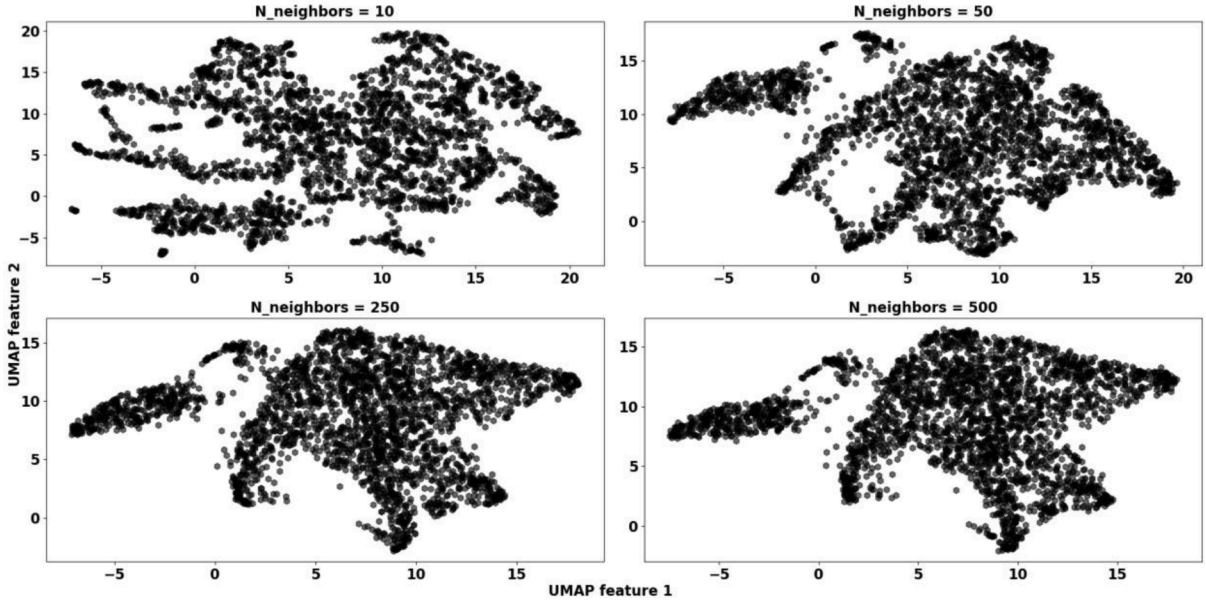


Fig. 7. Low dimensional representation of the original data by the UMAP algorithm with different n_neighbour's values

dimensional energy data. It is selected for further evaluation in this research to extract information from smart meters and building use characteristic data.

3.4. High variation regions identification

After selecting the high dimensional reduction method, the original data is mapped to low dimensional representation to understand the placement of data points by the UMAP algorithm. The n-neighbours and minimum distance parameters were set to 50 and 0.5 respectively enabling similar points to be close to each other with a clear distinction. The cluster representation is not done by an algorithm but with the original variable data of the high dimensional space that will lead to a more realistic selection. The area of interest is selected based on the higher values in a specific region and specified as a cluster. Fig. 9 depicts the mapping of the UMAP algorithm. The colour representation shows values from low to very high. The target variables are only evaluated for the related features of this study that are looking into the high self-production and consumption areas concerning building characteristics by eliminating the "Smart meter percentage" and the "Low tariff percentage" features. The regions with higher self-production and consumption are placed closer to each other by the UMAP algorithm as illustrated with cluster 1 and cluster 2. The higher self-production is coming from regions with higher consumption and a smaller number of active connections. It is understandable as most of the self-production comes from neighbourhoods with a low concentration of houses instead of city centres or buildings with apartments.

On the other hand, the higher consumption areas in cluster 2 are linked with a higher concentration of non-residential buildings represented by cluster 5. The neighbourhood with more non-residential buildings has a higher consumption even though the highest number of houses are in the region of cluster 6. This shows that compared to the

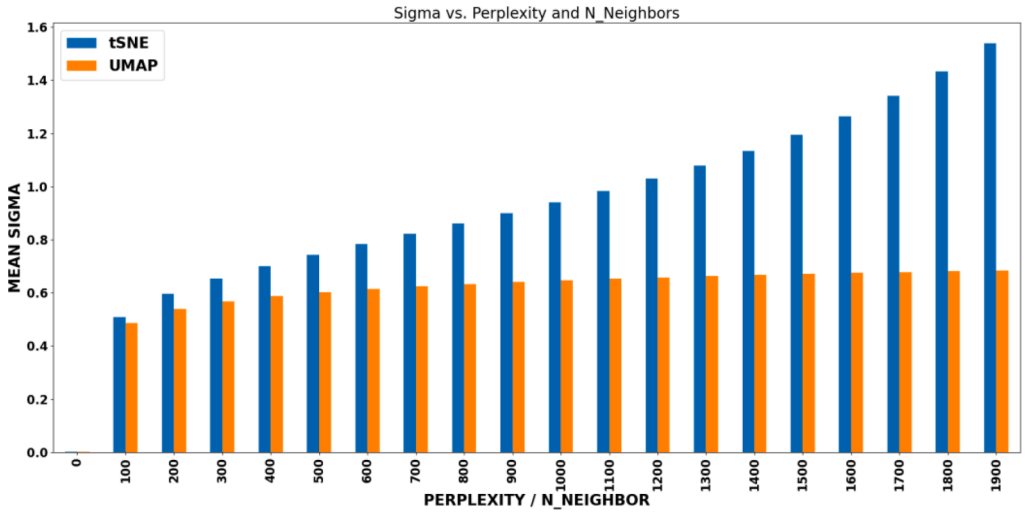


Fig. 8. Mean sigma visualisation with changing perplexity and n_neighbour's values

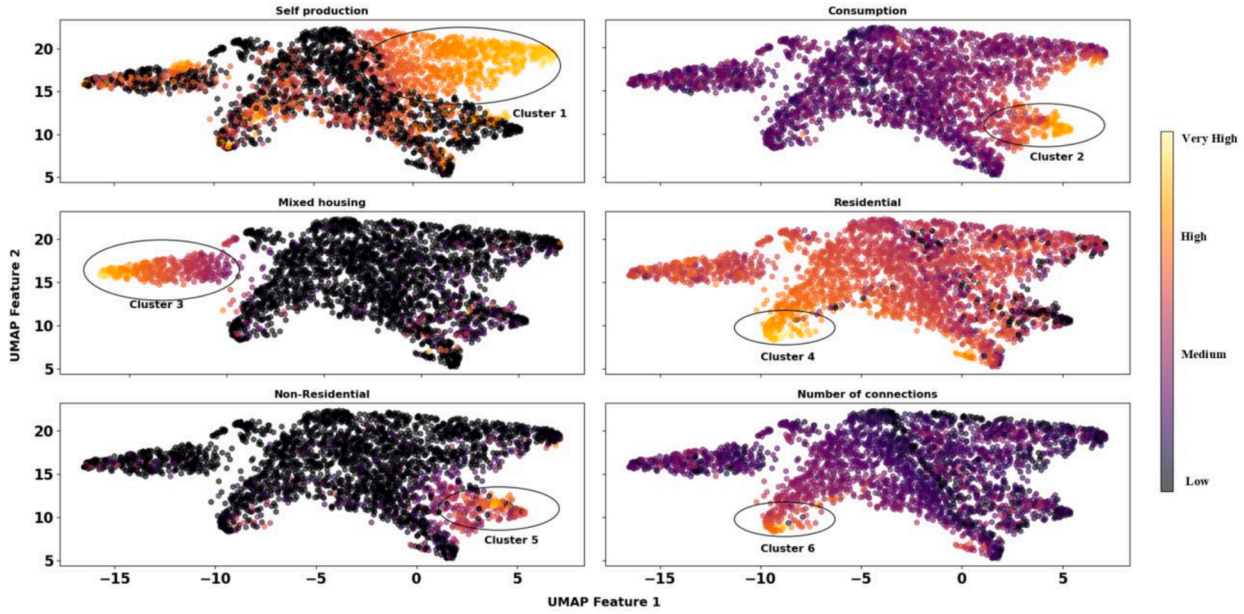


Fig. 9. Low dimensional visualisation using original data colour mapping by each feature to understand the clustered points by UMAP algorithm

number of residential buildings, non-residential buildings are more energy intensive. Additionally, the region of cluster 6 which represents areas with the most households has the lowest consumption and self-production. These neighbourhoods are mostly based in the city centre with a large number of apartment buildings and small houses. The UMAP algorithm also placed the high concentration of mixed buildings in the region of cluster 3. They have a higher self-production in some areas and an overall low to medium consumption.

The UMAP algorithm provides a quantitative assessment for the building energy data and identifies smaller focus areas from a bigger dataset that can help in better understanding the underlying aspects of these regions.

3.5. *Feature importance of using the MIC algorithm*

The MIC algorithm is applied here to understand the feature correlation of the reduced data with the input features. The algorithm is particularly applied to the high consumption and self-production regions for cluster 1 and cluster 2. The main goal of this was to investigate the importance of features for high consumption and self-production from the standpoint of building characteristics. Cluster 1 is based on the points with very high self-production based on the low dimensional space. Fig. 10 displays the MIC values for cluster 1. The Fig. provides the relationship between the original data variables with the UMAP generated target variables for both features. The feature importance is sorted based on UMAP feature 1 as it correlates more realistically with the high-generation neighbourhoods and confirms the UMAP data points placement. The self-production and residential building type parameters have the highest relationship which suggests that these two features are assigned higher weights in data points placement by the UMAP algorithm. Compared to the residential building concentration, the non-residential and mixed housing type showcase lower scores. This validates the previous section's initial assumption that the greater intrusion of self-production is primarily coming from residential buildings. From the MIC score of Cluster 2 in Fig. 11, it can be observed that the net consumption is higher in regions with more non-residential buildings. Furthermore, the residential building concentration does not affect the overall consumption compared to non-residential buildings, which is

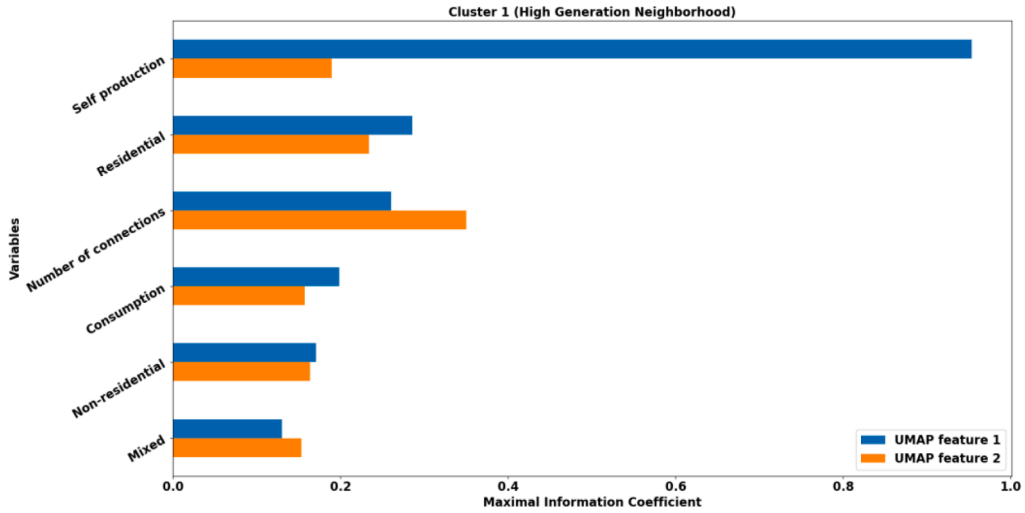


Fig. 10. MIC for feature importance of high self-production neighbourhoods in cluster 1

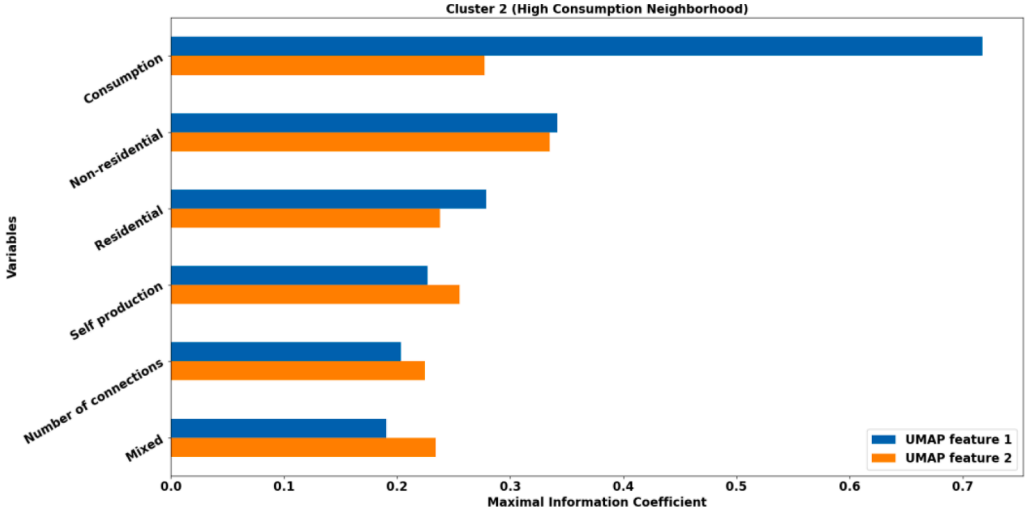


Fig. 11. MIC for feature importance of high consumption neighbourhoods in cluster 2

why the MIC score is low for residential buildings. This also proves that the MIC scores accurately identified the most relevant features of this cluster that were visible with the visualisation in Section 3.4.

In conclusion, the MIC method can help in understanding the relevance of a specific parameter for understanding regions of interest (clusters). It can provide the logic behind the mapping of dimensionality reduction. This process can be repeated for all the other regions, or the overall data placement based on the required outcome.

4. Discussion

A detailed understanding of smart cities energy data concerning the building characteristics using high-dimensional techniques is essential to better grasp the overall changes happening in the energy system infrastructure. Although the energy system still relies on major production sites, energy consumers now have more opportunities to become a prosumer. Many bottom-up efforts are being established in which people and small business owners attempt to become prosumers or even producers themselves, demonstrating this involvement (Hoppe et al., 2015). Nevertheless, the existing energy policies do not consider these bottom-up adaptations. It is necessary to develop new energy transition policies that can account for the effects of these developments. As a result, the electric utilities can improve their strategies and tailor packages to specific regions to decrease consumption. Whereas energy administrators can optimize and direct their focus on the identified regions to obtain their CO_2 emission reduction goals. However, despite the importance of these developments, it is difficult to access the data from buildings for a bottom-up approach and extract any meaningful information.

The aim of this research was to provide methodological contribution to the existing literature on urban energy analysis and planning from three perspectives as described in Section 1.2. First, this work presented a framework to evaluate the relationship between the energy consumption profile with the building characteristics. Second, it explored two non-linear dimensionality reduction techniques (UMAP and T-SNE) from the perspective of local and global preservation that is seldomly used in the literature. Third the framework provide logic behind the mapping of dimensionality reduction to confirm the validity of the obtained results using MIC algorithm.

According to the results obtained, the UMAP algorithm has comparably consistent visualisation and was able to reach an optimum structure for a high-dimensional energy dataset very quickly. In the t-SNE algorithm, the data points in low dimension representation are moving around as the t-SNE algorithm cannot find a global optimum for the value of sigma. This is a drawback of the t-SNE that leads to some false intuitions. Another cause for this could also be due to the t-SNE

method's random initialization or its cost function, as described in Section 2.3.3. UMAP uses a Laplacian eigenmaps representation for initialisation. That is why the UMAP algorithm retains the same structure and finds the global optimum easily. Another advantage of UMAP over t-SNE is the degree of closeness of points that can be controlled by a minimum distance hyperparameter. Even with a higher value of n_neighbour, it is possible to move the points for a clear representation with this parameter. In this research, this n-neighbours and minimum distance parameters were set to 50 and 0.5 respectively. The n-neighbours value was selected based on an exhaustive grid search to find the optimum value of the algorithm convergence. The minimum distance value was selected by visual inspection based on the data points separation.

Another important outcome of this research was the extraction of high consumption and self-production regions concerning building characteristics to help grid operators and energy planners in identifying areas of focus to achieve their CO_2 emission goals. The energy self-production profile can increase in areas with more residential buildings and can be hazardous for the grid side. There are several options for dealing with this problem. One possibility is to store the excess self-production in storage, either in electrical or thermal storage in the neighbourhoods with a higher number of residential buildings. Another possibility is to curtail the excess self-production in high solar irradiation periods in specific areas which will lead to some losses in the total energy yield (Li et al., 2020). The inclusion of the MIC algorithm also provides a technique for the validation of the obtained results in the low dimensional space with the original data.

The proposed framework also provides certain advantages over other techniques discussed in the literature review section. For example, the framework in this research can be applied on (Zarco-Periñán et al., 2021) to study the influence of population density on energy consumption. It is not required to aggregate all the city data into groups. The framework can cluster the cities based on consumption and population density and can also provide more details of the underlying causes if more parameters are included. The authors in (Kontokosta, 2015) investigated the causes of commercial building energy consumption across buildings, systems, geographic, and occupancy factors on actual building energy usage data. They visualised the relationship individually between two parameters using scatter plots. However, they were not able to visually explore all the parameters together and evaluate their nonlinear relationships. The current framework proposed in this paper has the advantage over simple visual analysis that it can combine all the information and present it as a global picture.

The trade-off between local and global structure preservation has always been a source of discussion for these strategies because they can only address one or the other. Our main objective in this work is to comprehend which parts of dimensionality reduction techniques are crucial for maintaining both local and global structures. It is difficult to extract information from high-dimensional data without a true understanding of the selection of the algorithms and their empirical impact on the low-dimensional space they produce. The analysis in the research reveals that the selection of an algorithm is critical for achieving the objective of global structure preservation and a post mining technique is necessary to confirm the validity of the obtained results.

4.1. Limitation of the proposed framework

The framework presented in this study was applied to the aggregated values of the neighbourhood's level. It can be implemented on individual building levels with information about different components and levels. However, the representation is limited to only aggregated information in a single row per class (building or neighbourhood). This limits the capability of the framework to only work on a unique temporal basis.

5. Conclusion

This work proposes a dimensionality reduction framework for the application of energy data analysis from neighbourhoods and cities. Data from 72000 households are analysed in a bottom-up approach to extract information and identify the underlying causes for the high consumption and self-production using dimension reduction techniques. Efforts are made to compare the performance of current non-linear dimensionality reduction methods and evaluate their performance in preserving the original data structure in reduced dimensional space. Furthermore, instead of clustering, colour mapping is used to identify the real underlying structure of the data. Finally, for the feature importance verification, the MIC method is proposed to detect linear and non-linear relationships between original and target variables. This provides a detailed perspective of the most relevant parameters, allowing for a more in-depth investigation to better comprehend the high dimensional data complexity.

Two high dimensional reduction techniques (t-SNE and UMAP) were explored and evaluated for visual analysis of building characteristics and energy. The techniques were cross-compared based on their strengths and limitations to best characterize low-dimensional behaviour. The results showed the UMAP's algorithm mean sigma quickly approaches a threshold of 0.6 and does not change with the increasing n_neighbor values. Whereas the t-SNE's mean sigma value increases continuously with the increasing perplexity value, implying that t-SNE is significantly more sensitive to the perplexity parameter. The comparison determined that the UMAP algorithm can have more consistent results by extracting similar representation in the low dimensional space from the original variables. The UMAP algorithm was selected as the main dimensionality reduction technique in this framework.

Based on the obtained results from the UMAP algorithm, the visual analysis indicated that self-production and consumption are linked with the amount of building use characteristics in a neighbourhood. As a growing number of residential buildings will install solar panels to reduce their energy expenses, the flow back to the grid will increase. Further research along these lines could help determine which alternative, or a mix of solutions, is the best fit for unlocking much-needed grid capacity. Overall, the proposed framework can enable stakeholders to identify areas of interest with underlying causes more accurately from high-dimensional energy datasets. An application of the proposed methodology can promote sustainable development and bring benefits to the community since understanding the energy profile at the neighbourhood and city level is essential to reduce the impact of climate change.

In the future perspective, other parameters from the building domain can be included in the analysis to explore the overall impact from the aspect of technological and demographical change or building characteristics such as the integration of EVs, the electrification of heat pumps, built years etc. Efforts are made with the help of the city to obtain data for multiple years and do a cross comparison between the changing energy profile of the city.

Another future possibility is to compare the changes in the energy profile based on incentives given by the city for the insulation of the buildings in a specific region. The framework has the potential to be widely used in the applications of urban energy analysis mainly depending on the availability of the data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The authors do not have permission to share data.

Acknowledgements

This study was funded by the Dutch Research Council (NWO). The building energy consumption data used in the study are supplied by the energy company, Netherlands and are only available in the form of tables and graphs presented in the study because of the restrictions on the use by third parties.

References

Cajot, S., Peter, M., Bahu, J.-M., Guignet, F., Koch, A., & Maréchal, F. (2017). Obstacles in energy planning at the urban scale. *Sustainable Cities and Society, 30*, 223–236. https://doi.org/10.1016/j.scs.2017.02.003. Apr.

Madlener, R., & Sunak, Y. (2011). Impacts of urbanization on urban structures and energy demand: What can we learn for urban energy planning and urbanization management? *Sustainable Cities and Society, 1*(1), 45–53. https://doi.org/10.1016/j.scs.2010.08.006. Feb.

Zhang, M., Zhang, X., Guo, S., Xu, X., Chen, J., & Wang, W. (2021). Urban micro-climate prediction through long short-term memory network with long-term monitoring for on-site building energy estimation. *Sustainable Cities and Society, 74*, Article 103227. https://doi.org/10.1016/j.scs.2021.103227. Nov.

Damsgaard, N., Helbrink, J., Papaefthymiou, G., Grave, K., Giordano, V., & Gentili, P. (2015). *Study on the effective integration of distributed energy resources for providing flexibility to the electricity system* https://doi.org/10.13140/RG.2.2.35386.39360. Report to the European commission.

Park, J. Y., Wilson, E., Parker, A., & Nagy, Z. (2020). The good, the bad, and the ugly: Data-driven load profile discord identification in a large building portfolio. *Energy and Buildings, 215*, Article 109892. https://doi.org/10.1016/j.enbuild.2020.109892. May.

Van Aubel, P., & Poll, E. (2019). Smart metering in the Netherlands: What, how, and why. *International Journal of Electrical Power & Energy Systems, 109*, 719–725. https://doi.org/10.1016/j.ijepes.2019.01.001. Jul.

The Netherlands 2020 "The Netherlands 2020 - Energy Policy Review," p. 258.

Wang, E., Shen, Z., & Grosskopf, K. (2014). Benchmarking energy performance of building envelopes through a selective residual-clustering approach using high dimensional dataset. *Energy and Buildings, 75*, 10–22. https://doi.org/10.1016/j.enbuild.2013.12.055. Jun.

Chen, Z., Freihaut, J., Lin, B., & Wang, C. D. (2018). Inverse energy model development via high-dimensional data analysis and sub-metering priority in building data monitoring. *Energy and Buildings, 172*, 116–124. https://doi.org/10.1016/j.enbuild.2018.04.061. Aug.

Sun, Y., Haghighat, F., & Fung, B. C. M. (2020). A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy and Buildings, 221*, Article 110022. https://doi.org/10.1016/j.enbuild.2020.110022. Aug.

Khan, W., Walker, S., & Zeiler, W. (2022). Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach. *Energy, 240*, Article 122812. https://doi.org/10.1016/j.energy.2021.122812. Feb.

Barber, K. A., & Krarti, M. (2022). A review of optimization based tools for design and control of building energy systems. *Renewable and Sustainable Energy Reviews, 160*, Article 112359. https://doi.org/10.1016/j.rser.2022.112359. May.

Wang, Z., & Hong, T. (2020). Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy, 269*, Article 115036. https://doi.org/10.1016/j.apenergy.2020.115036. Jul.

Lee, H., & Heo, Y. (2022). Simplified data-driven models for model predictive control of residential buildings. *Energy and Buildings, 265*, Article 112067. https://doi.org/10.1016/j.enbuild.2022.112067. Jun.

Wei, Y., et al. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews, 82*, 1027–1047. https://doi.org/10.1016/j.rser.2017.09.108. Feb.

Reshef, Y. A., Reshef, D. N., Finucane, H. K., Sabeti, P. C., & Mitzenmacher, M. (2011). *Measuring dependence powerfully and equitably*, 63. https://doi.org/10.1214/19-STS719

Rueda, L., Agbossou, K., Cardenas, A., Henao, N., & Kelouwani, S. (2020). A comprehensive review of approaches to building occupancy detection. *Building and Environment, 180*, Article 106966. https://doi.org/10.1016/j.buildenv.2020.106966. Aug.

Ngarambe, J., Yun, G. Y., & Santamouris, M. (2020). The use of artificial intelligence (AI) methods in the prediction of thermal comfort in buildings: energy implications of AI-based thermal comfort controls. *Energy and Buildings, 211*, Article 109807. https://doi.org/10.1016/j.enbuild.2020.109807. Mar.

Zhao, Y., Zhang, C., Zhang, Y., Wang, Z., & Li, J. (2020). A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. *Energy and Built Environment, 1*(2), 149–164. https://doi.org/10.1016/j.enbenv.2019.11.003. Apr.

Hong, T., Chen, Y., Luo, X., Luo, N., & Lee, S. H. (2020). Ten questions on urban building energy modeling. *Building and Environment, 168*, Article 106508. https://doi.org/10.1016/j.buildenv.2019.106508. Jan.

Torabi Moghadam, S., Delmastro, C., Corgnati, S. P., & Lombardi, P. (2017). Urban energy planning procedure for sustainable development in the built environment: A review of available spatial approaches. *Journal of Cleaner Production, 165*, 811–827. https://doi.org/10.1016/j.jclepro.2017.07.142. Nov.

Kontokosta, C. E., & Tull, C. (2017). A data-driven predictive model of city-scale energy use in buildings. *Applied Energy, 197*, 303–317. https://doi.org/10.1016/j.apenergy.2017.04.005. Jul.

Brownsword, R. A., Fleming, P. D., Powell, J. C., & Pearsall, N. (2005). Sustainable cities - modelling urban energy supply and demand. *Applied Energy, 82*(2), 167–180.

Heiple, S., & Sailor, D. (2008). Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles. *Energy and Buildings, 40*, 1426–1436. https://doi.org/10.1016/j.enbuild.2008.01.005. Dec.

Aranda, J., Zabalza, I., Llera-Sastresa, E., Scarpellini, S., & Alcalde, A. (2018). Building energy assessment and computer simulation applied to social housing in Spain. *Buildings, 8*, 11. https://doi.org/10.3390/buildings8010011. Jan.

Kontokosta, C. E. (2015). A market-specific methodology for a commercial building energy performance index. *Journal of Real Estate Finance and Economics, 51*(2), 288–316. https://doi.org/10.1007/s11146-014-9481-0. Aug.

Zarco-Periñán, P. J., Zarco-Soto, I. M., & Zarco-Soto, F. J. (2021). Influence of the population density of cities on energy consumption of their households. *Sustainability, 13*(14). https://doi.org/10.3390/su13147542. Art. no. 14Jan.

Hong, T., Chou, S. K., & Bong, T. Y. (2000). Building simulation: an overview of developments and information sources. *Building and Environment, 35*(4), 347–361. https://doi.org/10.1016/S0360-1323(99)00023-2. May.

Ferrando, M., Causone, F., Hong, T., & Chen, Y. (2020). Urban building energy modeling (UBEM) tools: A state-of-the-art review of bottom-up physics-based approaches. *Sustainable Cities and Society, 62*, Article 102408. https://doi.org/10.1016/j.scs.2020.102408. Nov.

ABC, "Dimensionality reduction for visualizing single-cell data using UMAP | Nature Biotechnology." https://www.nature.com/articles/nbt.4314 (accessed Apr. 20, 2022).

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9*(86), 2579–2605.

L. McInnes, J. Healy, and J. Melville, "UMAP: uniform manifold approximation and projection for dimension reduction," *arXiv:1802.03426 [cs, stat]*, Sep. 2020, Accessed: Apr. 01, 2022. [Online]. Available: http://arxiv.org/abs/1802.03426.

ABC, "Concurrent time-series selections using deep learning and dimension reduction - ScienceDirect." https://www.sciencedirect.com/science/article/pii/S0950705121007693 (accessed Jun. 24, 2022).

C. B. voor de Statistiek, "Centraal Bureau voor de Statistiek," Centraal Bureau voor de Statistiek. https://www.cbs.nl/ (accessed Apr. 19, 2022).

N. Lewin-Koh, "Hexagon binning: An overview." Jan. 08, 2021. [Online]. Available: https://cran.r-project.org/web/packages/hexbin/vignettes/hexagon_binning.pdf.

Burdziej, J. (2019). Using hexagonal grids and network analysis for spatial accessibility assessment in urban environments - A case study of public amenities in Toruń. *Miscellanea Geographica, 23*. https://doi.org/10.2478/mgrsd-2018-0037. Jan.

ABC, "Introduction guide to contextily — contextily 1.1.0 documentation." https://contextily.readthedocs.io/en/latest/intro_guide.html (accessed Apr. 20, 2022).

Berman, J. J. (2018). 11 - Indispensable tips for fast and simple big data analysis. In J. J. Berman (Ed.), *Principles and practice of big data* (2nd Ed., pp. 231–257). Academic Press. https://doi.org/10.1016/B978-0-12-815609-4.00011-X.

L. S. Nelson, "The anderson-darling test for normality," 1998, doi: 10.1080/00224065.1998.11979858.

Jäntschi, L., & Bolboacă, S. D. (2018). Computation of probability associated with anderson–darling statistic. *Mathematics, 6*(6). https://doi.org/10.3390/math6060088. Art. no. 6Jun.

ABC, "scipy.stats.anderson — SciPy v1.8.1 Manual." https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.anderson.html (accessed Jun. 07, 2022).

Vachharajani, B., & Pandya, D. (2022). Dimension reduction techniques: Current status and perspectives. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.12.549. Jan.

El Bouchefry, K., & de Souza, R. S. (2020). Chapter 12 - Learning in big data: Introduction to machine learning. In P. Škoda, & F. Adam (Eds.), *Knowledge discovery in big data from astronomy and earth observation* (pp. 225–249). Elsevier. https://doi.org/10.1016/B978-0-12-819154-5.00023-0.

Jiale, Y., & Ying, Z. (2020). Visualization method of sound effect retrieval based on UMAP. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (pp. 2216–2220). https://doi.org/10.1109/ITNEC48623.2020.9085193. Jun.vol. 1.

Demidova, L., & Stepanov, M. (2020). Data analysis using the nonlinear dimension reduction algorithms. In *2020 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)* (pp. 211–216). https://doi.org/10.1109/SUMMA50634.2020.9280727. Nov.

Pal, K., & Sharma, M. (2020). Performance evaluation of non-linear techniques UMAP and t-SNE for data in higher dimensional topological space. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (pp. 1106–1110). https://doi.org/10.1109/I-SMAC49090.2020.9243502. Oct.

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications, 10*(1). https://doi.org/10.1038/s41467-019-13056-x. Art. no. 1Nov.

Pouyet, E., Rohani, N., Katsaggelos, A., Cossairt, O., & Walton, M. (2018). Innovative data reduction and visualization strategy for hyperspectral imaging datasets using t-SNE approach. *Pure and Applied Chemistry, 90*. https://doi.org/10.1515/pac-2017-0907. Jan.

Kingman, J. F. C. (1970). Information theory and statistics. By Solomon Kullback. Pp. 399. 28s. 6d. 1968. (Dover.). *The Mathematical Gazette, 54*(387), 90. https://doi.org/10.2307/3613211. –90Feb.

ABC, "2.2. Manifold learning," scikit-learn. https://scikit-learn/stable/modules/manifold.html (accessed Aug. 03, 2022).
van der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research, 15*(93), 3221–3245.
ABC, "tSNE vs. UMAP: Global Structure. Why preservation of global structure… | by Nikolay Oskolkov | Towards Data Science." https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17 (accessed Apr. 20, 2022).
Z. Yang, Y. Chen, and J. Corander, "T-SNE is not optimized to reveal clusters in data," *arXiv:2110.02573 [cs, stat]*, Oct. 2021, Accessed: Apr. 20, 2022. [Online]. Available: http://arxiv.org/abs/2110.02573.
Nguyen, L. H., & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLOS Computational Biology, 15*(6), Article e1006907. https://doi.org/10.1371/journal.pcbi.1006907. Jun.
Song, L., Langfelder, P., & Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics, 13* (1), 328. https://doi.org/10.1186/1471-2105-13-328. Dec.
Reshef, D. N., et al. (2011). Detecting novel associations in large datasets. *Science, 334* (6062), 1518–1524. https://doi.org/10.1126/science.1205438. Dec.
ABC. (2022). *sklearn.manifold.TSNE*. scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.manifold.TSNE.html accessed Apr. 20.
L. McInnes, "lmcinnes/umap." Apr. 20, 2022. Accessed: Apr. 20, 2022. [Online]. Available: https://github.com/lmcinnes/umap.
N. Oskolkov, "NikolayOskolkov/tSNE_vs_UMAP_GlobalStructure." Apr. 15, 2022. Accessed: Apr. 26, 2022. [Online]. Available: https://github.com/NikolayOskolkov/tSNE_vs_UMAP_GlobalStructure.
ABC, "MICtools." minepy - Maximal Information-based Nonparametric Exploration (MINE) in C and Python, Mar. 23, 2022. Accessed: Apr. 26, 2022. [Online]. Available: https://github.com/minepy/mictools.
Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., & Furlanello, C. (2013). minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics, 29*(3), 407–408. https://doi.org/10.1093/bioinformatics/bts707. Feb.
Hoppe, T., Graf, A., Warbroek, W. D. B., Lammers, I., & Lepping, I. (2015). Local governments supporting local energy initiatives: Lessons from the best practices of Saerbeck (Germany) and Lochem (The Netherlands). *Sustainability (Switzerland), 7* (2), 1900–1931. https://doi.org/10.3390/su7021900
Li, H. X., Zhang, Y., Edwards, D., & Hosseini, M. R. (2020). Improving the energy production of roof-top solar PV systems through roof design. *Building Simulation, 13* (2), 475–487. https://doi.org/10.1007/s12273-019-0585-6. Apr.