Energy & Buildings 238 (2021) 110815

Contents lists available at ScienceDirect

Energy & Buildings

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

Data-driven optimization of building layouts for energy efficiency

Andrew Sonta *, Thomas R. Dougherty, Rishee K. Jain

Urban Informatics Lab, Department of Civil and Environmental Engineering, Stanford University, 473 Via Ortega Rm 269B, Stanford, CA 94305, United States

ARTICLE INFO

Article history:
Received 28 October 2020
Revised 16 January 2021
Accepted 6 February 2021
Available online 11 February 2021

Keywords:
Occupant dynamics
Design optimization
Data-driven simulation
Energy efficiency
Machine learning

ABSTRACT

One of the primary driving factors in building energy performance is occupant behavioral dynamics. As a result, the layout of building occupant workstations is likely to influence energy consumption. In this paper, we introduce methods for relating lighting zone energy to zone-level occupant dynamics, simulating energy consumption of a lighting system based on this relationship, and optimizing the layouts of buildings. The optimization makes use of both a clustering-based approach and a genetic algorithm, and it aims to reduce energy consumption. We find in a case study that nonhomogeneous behavior (i.e., high diversity) among occupant schedules positively correlates with the energy consumption of a highly controllable lighting system. We additionally find through data-driven simulation that the naïve clustering-based optimization and the genetic algorithm (which makes use of the energy simulation engine) produce layouts that reduce energy consumption by roughly 5% compared to the existing layout of a real office space comprised of 151 occupants. Overall, this study demonstrates the merits of utilizing low-cost dynamic design of existing building layouts as a means to reduce energy usage. Our work provides an additional path to reach our sustainable energy goals in the built environment through new non-capital-intensive interventions.

1. Introduction

The energy performance of buildings is largely driven by the operation of their energy-intensive systems. In commercial office buildings, the most energy-intensive systems are those that provide comfortable thermal and visual environments (i.e., heating, cooling, and lighting systems). The operation of these systems critically depends on the subjective experience of building occupants when they are using a building's spaces. As a result, there is no need to heat, cool, or light spaces that are not used at a particular point in time. These thermal and lighting systems are often controlled by zone, and in the case that even one occupant enters a zone, the systems must typically service the entire zone. This shared feature of building system operation contributes heavily to inefficiency [1], but it also creates the opportunity to optimize the design and management of building spaces and save energy through the individualization of building spaces. In this work, we investigate the optimization of layouts in existing buildings in response to data on occupant use of space. We define the layout as the assignment of occupants to the existing workstations in the building. The flexibility in determining which occupants sit at which workstations forms the possible design space for layout optimization.

Let us consider a hypothetical example of an office building with 4 teams, 4 members per team, and 4 shared rooms/zones. Let us assume, for the purposes of this example, that this office building houses a global call center, whereby one member of each team must be present in the office to make calls at any given time. All 4 teams have decided to operate on the following schedule: the first team member works from 12 am to 6 am, the second from 6 am to 12 pm, the third from 12 pm to 6 pm, and the fourth from 6 pm to 12 am. If the office is arranged such that each team occupies its own room/zone, there will be one person in each room at all times. In other words, all 4 rooms will have exactly 1 occupant in the room at all times. We refer to this situation as an example of high *occupant diversity* within each building zone, where diversity is a term used to describe the differences in activities or schedules. As a result of this diversity, the heating, cooling, and lighting systems for all four rooms will operate at all times. We could instead arrange the layout such that each occupant shares a room with their fellow shift-workers (e.g., all 12 am to 6 am workers share a room). In this case, only one room, the occupied room, will need to be supplied with heating, cooling, and lighting throughout the day. The operation time of these systems would therefore be reduced by 75% compared to the first scenario.

* Corresponding author.
E-mail address: asonta@stanford.edu (A. Sonta).

https://doi.org/10.1016/j.enbuild.2021.110815
0378-7788/

While this is an extreme example, the underlying dynamics apply to all buildings with shared spaces. In reality, the complexities and subtleties of occupant schedules, particularly in large office buildings, make it difficult to discern shared patterns of behavior. The decisions of creating building layouts, therefore, typically involve functional or hierarchical structures of organizations. In this paper, we investigate the possibility of optimizing layouts based on occupants' use of spaces.

Past research into building energy efficiency has focused on both building design as well as building operation. A large body of work has focused recently on the impact of occupant behavior on energy-intensive operation of building systems. Furthermore, researchers have investigated the role of design optimization, including through an occupant-centric lens, in reducing building energy consumption. Below, we discuss key findings from previous research and motivate the work presented in this paper.

1.1. Building design and energy efficiency

It is well known that the design of buildings has a large impact on their future operation, including energy efficiency [2]. When considering energy in building design, architects and engineers generally consider physical building parameters including orientation, materiality, fenestration, and choice of heating, cooling, and lighting systems [3–5]. The design process also considers, in addition to these physical characteristics of individual components, the impacts of layout on occupant behavior through the architectural programming process. This programming is often driven by intended building use (e.g., meeting rooms or workspaces in an office building). While this design process is typically regarded in terms of our subjective experience of the building, past research has shown that these choices for building layouts can influence the energy consumption of the building [1]. For example, locating parts of a home that are more often used in the morning on the eastern side of the house could reduce the amount of heating, cooling, and lighting required for those spaces.

An advantage of considering the layout of the building in the context of energy efficiency is that layouts are often flexible, especially compared to other design considerations such as orientation and materiality. This is especially important due to the fact that our use of buildings evolves over time. Buildings originally intended for one purpose are often repurposed to suit changing needs, leading to different patterns of space utilization and rendering some physical design considerations obsolete. Moreover, as buildings lifespans are on the order of 40–100 years [6], it is expected that the majority of energy consumption from the building sector will come from existing buildings rather than new construction for many upcoming years—years that are critical to our sustainable energy goals [2]. This importance of the existing building stock in addressing energy challenges suggests the need for new and innovative ways to reconsider the energy-intensive attributes of existing building design. New design methods that focus on building layouts are a promising means of addressing these challenges.

1.2. The role of the occupant in energy efficiency

As discussed above, building layouts are often driven by intended use. Research has shown that the actual use of spaces by building occupants, while extremely impactful to building energy consumption, is very difficult to model and understand [7–9]. For example, a seminal study showed that changes in occupant behavior can cause two-to-one discrepancies in actual versus expected energy consumption [10]. Both occupant actions (e.g., interactions with building systems through thermostats, windows, etc.) as well as their passive use of space are important to energy consumption [11]. This latter notion, passive space use, largely drives energy consumption in buildings that have systems that are able to respond (e.g., turn off or reduce service) based on occupancy information. These responsive systems are increasing in prevalence as energy codes and regulations are driving wider adoption of energy-efficient technologies in the building sector [12].

1.3. Optimizing buildings for energy efficiency

As both design and occupant behavior are key factors that affect the energy efficiency of buildings, researchers have sought to leverage optimization of building design and operation as a tool to improve energy performance. These optimization approaches have generally focused either on building design pre-construction or operation of heating, cooling, and lighting systems post-construction. Design optimization is typically considered in the context of new construction [3]. Such research has largely focused on physical building parameters in the early stages of design [3,13]. However, because of the *coupled* effects of occupant behavior and building design on building performance, researchers have noted that occupant-related uncertainty can hinder confidence in early-stage design decisions supported by such optimizations [14]. These coupled dynamics have driven researchers to focus more on the operational phase of building—and particularly the importance of building occupants—as a means of improving energy efficiency [12].

Researchers have therefore developed data-driven tools that optimize control of energy-intensive systems in existing buildings, finding that optimizing control of systems through an occupant lens can enable large reductions in energy consumption, ranging from 15 to 70% depending on the types of systems analyzed [15–17]. Recently, researchers have found that an important component of such control optimization strategies is the explicit consideration of occupant comfort, which improves the subjective experience of the occupants but can reduce the energy savings possible [18]. This research shows the promise of introducing sustainable energy savings in existing buildings by controlling building systems optimally. However, this research also considers the dynamics of occupant behavior as given. To take full advantage of controllable building systems, there remains the opportunity not only to optimize how systems respond to building space use, but also to optimize how people themselves use spaces—an aspect we consider in this paper through the building layout.

The previous research discussed in this section has shown that we can optimize building designs before construction and optimize building controls after construction. However, the dynamic behavior of building occupants—behavior that enables research into building controls—can also enable new research into building design features that remain flexible once the building is erected. As discussed above, a focus on building layouts may offer a means for addressing energy gaps in design through a naturally occupant-focused lens. A key research gap, therefore, is the integration of design optimization with important characteristics of dynamic occupant behavior. While optimization of physical building parameters before construction will remain a promising area of inquiry, there is a need to be able to dynamically examine aspects of building design once its key operational parameter, the building occupant, enters the picture.

Some emerging research at the intersection of building design, operation, and occupant experience concerns optimal utilization of designed spaces. For example, the paradigm of activity-based workspaces (ABW) seeks to provide occupants with different spaces for different tasks, as opposed to the more traditional paradigm of individual desk assignment [19]. Recent work has shown that computational approaches can leverage the ABW paradigm

to optimize energy and space use efficiency [20]. Along similar lines, researchers have considered the notion of "green scheduling," whereby certain activities are scheduled for particular building zones or at particular times of the day in an effort to reduce building system operation time, for example in the case of scheduling course timetables in university buildings [21]. These approaches explicitly consider the interaction between occupant experience and building energy consumption, working toward an occupant-centric building design and operation paradigm. Research in these areas tends to focus on analysis of existing building layouts, but there also remains the opportunity to develop methods for reimagining the layouts of existing buildings.

Recent research has investigated such dynamic optimization of building layout designs in terms of organizational structure and performance [22]. For example, Lather et al. [23] considered the layout of hospitals based on organizational adjacencies among healthcare departments, using a graph theoretical approach to develop optimal layouts. Similarly, Lee et al. [24] defined a building's operating efficiency in terms of occupants' estimated walking time as they carried out different activities, leveraging ant colony optimization to develop building floor plans. However, the direct optimization of building layouts to address the energy-efficient operation of building systems remains an area for continual research.

In this paper, our overarching research question is whether building layouts can be optimized to reduce energy consumption of energy-intensive systems by leveraging sensor data on occupant activities. We first describe our previously introduced methods for abstracting time series plug load data to states of occupant behavior. We then discuss our methods of analyzing zone-level diversity in occupant schedules, optimizing layouts, and simulating the impacts of adopting an optimized layout. We introduce a real-world case study, where we apply our methods to a floor of an office building with 151 occupants.

2. Methodology and data collection

In this section, we describe our overall methodology for developing a framework that enables the optimization of building layouts—defined as the allocation of occupants to existing workstations—for energy savings (Fig. 1). We first leveraged ambient sensor data collected from plug load energy sensors at the individual desk level to describe occupants' use of space over time. We term this description the individual's *occupant schedule* (Section 2.1). We define a distance metric that can be used to describe the *zone diversity* in occupant schedules over several individuals (Section 2.2). This distance metric creates the ability to cluster occupants spatially (i.e., create new layouts) in an effort to reduce this zone diversity and therefore building energy consumption (Section 2.3). Due to the extremely high dimensionality of the possible solution space, we also introduce a genetic algorithm for creating occupant layouts based on expected energy consumption (Section 2.4). For evaluation of our layout optimization algorithms, we introduce a data-driven surrogate model for simulating energy consumption of a building's lighting system based on occupant schedules (Section 2.5). We describe the dataset used for analysis in Section 2.6.

2.1. Data collection and preprocessing: learning occupant schedules from plug-load sensor data

Our analysis approach makes use of time series sensor data collected from plug load sensors installed at individual desks in commercial office buildings—data that provides insight into patterns of space use. We note that other sources of individualized sensor data that can describe individual patterns of space use could be adapted as the underlying data. We define the time series energy data as $\mathbf{X}_{i,d}$ where i is the occupant index (for all occupants 1,. . .,I) and d is the day index (for all days 1,. . .,D). The total number of time steps is $D \times T$, where D is the number of days and T is the number of time steps during the day (i.e., if we collect data at 15-minute intervals, $T = 96$). Each entry in the data matrix, $\mathbf{X}_{i,d}$ can be represented as a vector $\{x_1,. . .,x_T\}$. We leveraged the method introduced in Sonta et al. [25] to map the raw data to abstracted states of occupant activities: $\mathbf{X}_{i,d} \rightarrow \mathbf{S}_{i,d}$. This mapping, based on variational Bayesian inference with a Gaussian Mixture Model (VB-GMM), was used to cluster the time series data into discrete states. For each observation x_t, we introduced a latent variable, $\mathbf{z}_i$ comprising a 1-of-K binary vector with elements $z_{t,k}$ for $k = 1,. . .,K$, where K is the possible number of components that can be used to cluster the data. Given the set of weights for each component k, which we refer to as $\boldsymbol{\phi}$, we can write the conditional distribution of $\mathbf{Z}$ as follows:

$$p(\mathbf{Z}|\boldsymbol{\phi}) = \prod_{t=1}^{T}\prod_{k=1}^{K}\phi_k^{z_{t,k}}$$

The conditional distribution of the observed plug load data, $\mathbf{X}$, can therefore be written as follows, given the latent variables and component weights:

$$p(\mathbf{X}|\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\Lambda}) = \prod_{t=1}^{T}\prod_{k=1}^{K}N\left(x_t|\mu_k,\Lambda_k^{-1}\right)^{z_{tk}}$$

where $\boldsymbol{\mu}$ is the set of component means and $\boldsymbol{\Lambda}$ is the set of component precisions (defined as the inverse of the standard deviations). Following standard Bayesian statistical practices, we introduced a Dirichlet distribution over the mixing coefficients and a Gaussian-Wishart prior over the mean and precision of each component. One of the key outputs of fitting this model is the number of components in $\boldsymbol{\phi}$ that are non-zero. The resulting non-zero distributions were then used to cluster the data.

As discussed in Sonta et al. [25], we used a two-step process for finding the number of components. If the initial clustering of the data resulted in two components, we separated out the higher-energy data and re-ran the clustering algorithm. Our rationale for doing this is based on our domain knowledge of occupant behavior and plug load data—the higher energy data has high variability and is likely to represent multiple states of activity. Consistent with

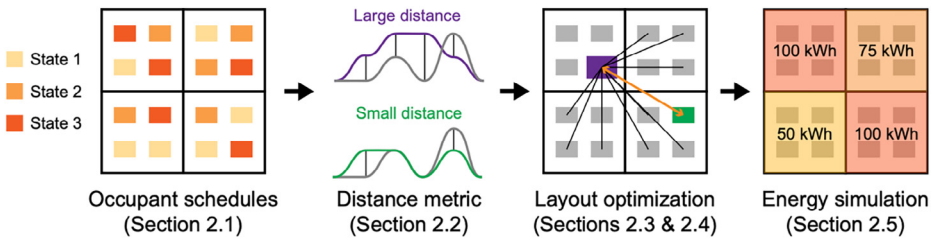

Fig. 1. Outline of methodology. Section 2.1 covers inference of occupant schedules from plug load data, 2.2 covers the zone diversity metric, 2.3 and 2.4 cover layout optimization schemes, and 2.5 covers the data-driven surrogate model for energy simulation.

previous results, the model output was most commonly two components for the initial clustering and two components for the secondary clustering. We therefore applied this two-step two-component clustering to the data in this study. The result was a mapping $\mathbf{X} \rightarrow \mathbf{S}$, where $s_{i,d}^{t} \in \{1, 2, 3\}$.

These three activity states, due to their construction, reflect intensity of energy use; we therefore refer to these states as low energy, medium energy, and high energy. This intensity of energy use creates the physical interpretation of each activity state. For example, higher energy use values map to high energy activity states, which correspond to occupants actively using their workstations. Similarly, a medium energy activity state is likely to signify that some equipment has entered a power-saving mode without fully turning off. Past work has shown that this mapping of the ambient plug load data to occupant activity states constitutes an occupancy sensing strategy at least as accurate as other state-of-the-art sensing strategies, such as infrared sensors [25,26]. An added advantage of the activity state strategy is that it offers additional information beyond presence/absence in that it describes the state of interaction with the workstation (e.g., a high energy activity state suggests full interaction with the workstation equipment). We refer to these individual time series of activity states as occupant schedules hereafter.

2.2. Representing zone diversity of occupant schedules using Euclidean distance

As discussed above, buildings provide energy-intensive services by zones, and the spatial efficiency of providing these services depends on occupant schedules. Therefore, a key question in understanding the operation of these systems, from the perspective of spatial efficiency, concerns the similarities or differences among the schedules of occupants within individual zones. We term these similarities and differences as the *diversity* in occupant schedules among occupants within a given space, and we operationalize this measure on time series data. Based on the work in Yang et al. [1], we can define this diversity as the distance between the vectors describing time series schedules for each occupant in the zone. A range of distance metrics could be used, including cosine similarity, Manhattan distance, Euclidean distance, etc. Following previous research practice [1], we used the Euclidean distance for this study but note that the specific distance choice did not have a large impact on the analysis.

If our schedule data is structured as above—$\mathbf{S}_{i,t}$, where i is the occupant index and t is the time index—we can compute Euclidean distances between the schedules for any two occupants. For example, the distance between occupant i and occupant j can be computed as follows:

$$d_{i,j} = \sqrt{\sum_{t=0}^{T} \left(\mathbf{S}_{i,t} - \mathbf{S}_{j,t}\right)^2}$$

Using this distance metric, we computed the distances between all occupants in a zone, forming a distance matrix. Normalizing this distance matrix by the total number of entries in the matrix (except the diagonal, since occupants' distance from themselves is 0), we have an average distance among all the occupant schedules within the zone, which we define as the *overall zone diversity*. With this metric, we compared the diversity of occupant schedules for individual building zones to the actual energy consumption of the building systems. We would expect higher zone diversity to correlate with higher energy consumption, as was shown using physics-based simulation in Yang et al. [1].

2.3. Optimizing layouts: Dimensionality reduction and occupant clustering

2.3.1. Dimensionality reduction using truncated singular value decomposition

Given occupant activity states and the notion of zone diversity in occupant dynamics, our next objective was to create optimal groupings of occupants in space—that is, to optimize a building's layout as a means to minimize energy use. Time series sensing generally produces many signals over time for each sensor deployed. In our case, activity states are generally reported on the scale of 15 min, creating 96 signals per occupant per day, or up to 35,000 signals per year. Computation of distances between vectors of this size suffers from the well-documented curse of dimensionality, whereby distance functions lose their usefulness as the dimension of the space increases [27]. We therefore employed a common dimensionality reduction process known as Singular Value Decomposition (SVD), a generalization of Primary Component Analysis, as a means to reduce the dimensionality of our data. We note that the zone diversity metric introduced above can be computed either for the unreduced data or the reduced data—the definition holds for both perspectives.

A common technique in recommendation algorithms and dimensionality reduction, SVD allows the reduction in size of the data while still capturing valuable features [28]. This dimensionality reduction is done by projecting the data with a set of orthogonal basis vectors representing the modes of variance in the system. These vectors are often referred to as the "concept space", as each vector represents some abstract concept which captures the variance. Truncating the lower variance orthogonal vectors before reconstruction yields a best approximation of the data in lower dimensional space, and we can select the number of dimensions depending on how much information we would like to retain during reconstruction. We applied SVD to a transposed version of our activity states: $\mathbf{M} = \mathbf{S}^{\mathrm{T}}$ (where $\mathbf{M}$ has $D \times T$ time rows and I occupant columns) Here, the number of rows is expected to be much larger than the number of columns. The decomposition is as follows:

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$$

Here, $\mathbf{U}$ contains the eigenvectors of $\mathbf{M}\mathbf{M}^{\mathrm{T}}$, $\mathbf{V}$ contains the eigenvectors of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$, and $\boldsymbol{\Sigma}$ is a diagonal matrix containing the singular values of $\mathbf{M}$ ($\boldsymbol{\Sigma}^2$ contains the eigenvalues of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$). The shapes of these matrices are determined by r = rank($\mathbf{M}$), where in our case, $r = I$. Therefore, $\mathbf{U}$ has the shape ($D \times T$, I). These eigenvector matrices can be thought of as a rigid transformation in high dimensional space, which aligns the data according to the variance of the data. Thus the primary axis after the rotation will be aligned with the axis of highest variance, the second axis will be aligned with the second highest variance, etc. In practice, the matrices $\mathbf{U}$ and $\mathbf{V}$ map the data to a concept space. The concept space is defined by the shape of the primary eigenvectors in the system, which typically will provide some kind of intuition as to what is driving the variance. In occupant schedules, a concept space might identify the time at which a person usually arrives at the office to be a valuable indicator, or when they take their lunch break.

Our purpose for using SVD was to project very high dimensional occupant behavioral space to a much lower dimensional representation, permitting a richer and faster clustering process. After the matrix was decomposed into $\mathbf{U}$, $\mathbf{V}$, and $\boldsymbol{\Sigma}$, the original data matrix $\mathbf{M}$ was projected into concept space via a rigid rotation from $\mathbf{U}$: $\mathbf{R} = \mathbf{U}^{\mathrm{T}}\mathbf{M}$. The result of this projection was the condensing of the state data into an $I \times I$ matrix $\mathbf{R}$, where each column, previously the length of the full time series, is now more densely represented,

and the different columns correspond to the different occupants. Because **U** is orthogonal, the removal of the least powerful vectors prior to multiplication with **M** yielded a projection of the data into the lower dimensional space with the least amount of lost value. For example, for a d-dimensional representation of the data, such that $d + d' = I$, the least significant d' eigenvectors/values in **U** are removed and the resulting representation of the full dataset $\mathbf{R}'$ will have the dimension $d \times I$. This approximates the replacement of the smallest eigenvalues with a 0 term, which becomes redundant in the reconstruction and can be truncated without losing value.

2.3.2. Stochastic constrained expectation maximization occupant clustering

With this relatively low-dimensional representation of our activity data **S**, data-driven clustering of the occupants according to their activity states becomes more feasible. Here, we introduce a novel clustering algorithm based on the data representation $\mathbf{R}'$. The objective of the clustering algorithm is to minimize the zone diversity metric from Section 2.2 for the occupants within a building zone. Our problem setting has the real-world constraint that each of the building zones retain their same size at the end of the clustering routine to preserve the same overall occupant spatial density, preventing the use of standard clustering algorithms such as k-means. The intuition behind our novel approach is to minimize the zone diversity metric by spatially swapping occupants with other occupants that reduce zone diversity. By doing so, we can expect to reduce lighting consumption according to the relationship we have established between these metrics.

The mechanics of the algorithm are depicted in Fig. 2. First, we chose, with replacement, a random occupant/desk, which is associated with a building zone. We note that building zones do not all need to be the same size as depicted in Fig. 2. We then simulated a "swap" between this occupant and all occupants in the other zones of the building. The resulting swap was the shift which had the greatest overall drop in zone diversity, which includes the null action of the occupant swapping with itself. We repeated this process until an iteration limit was reached. We note that alternative stopping criteria could be used, such stopping whenever the improvement after iterating is below a certain threshold. In our algorithm, however, the null action of the occupant swapping with itself would have the effect of no improvement. For ease of implementation, the iteration limit approach offers simplicity along with flexibility to handle such edge cases.

2.4. Optimizing layouts: Genetic algorithm

Our problem statement, to optimize building layout in order to reduce the energy consumption of building systems, has an extremely large solution space. If there are I occupants in a building and n possible zones to assign them to (of equal size $m = I/n$), then the number of possible assignments can be computed as follows:

$$\text{groups} = \frac{I!}{(m!)^n \times n!}$$

For example, with 50 occupants and 5 zones, the number of possible assignments is on the order of 10^{29}. In addition to the large solution space associated with our problem, the effects of reassigning occupants would be expected to be highly nonlinear. When optimizing in these circumstances, genetic algorithms have been shown to perform well [29–31]. We therefore implemented a custom genetic algorithm to assign occupants to desks and optimize building layouts, as described below.

Genetic algorithms belong to the class of evolutionary algorithms for optimization, originally inspired by the process of natural selection. The process begins by creating an initial set of design points—in our case, building layouts. Each building layout x in the initial population **P** was defined by the grouping of occupants to the zones of a building. A fitness function was used to evaluate the fitness of each design point $f(x)$. For this fitness function, we leveraged a data-driven surrogate simulation engine that can be used to predict building energy consumption based on occupant schedules and other time series information, as discussed below in Section 2.5. Once each design point was evaluated, a certain number of designs were selected to create a new generation of designs. In our case, we selected the **B** best performing layouts, and, in order to maintain diversity in the population, we also selected **R** random layouts. Among the best layouts and randomly selected layouts, two layouts were chosen at a time and recombined c times to form the layouts in the next generation. The first step in recombination was crossover, whereby for each desk location in each zone, the occupant selected to occupy that desk was a random selection of the two occupants in the original two designs. The next step was mutation, which occured for each new individual with probability m. If mutation did occur, a random desk in each zone was swapped with a random desk from a random other zone. Crossover is meant to preserve the high-performing features that exist in the best-performing layouts in the previous generation; mutation is meant to introduce randomness so that the algorithm does not get stuck in a local minimum. Once a completely new generation was created from the previous generation, through crossover and mutation, the process repeated for G generations. The parameters, therefore, that must be chosen to run the genetic algorithm are the fitness function f, the population size $|\mathbf{P}|$, the number of best performing layouts $|\mathbf{B}|$ and the number of randomly chosen layouts $|\mathbf{R}|$, the number of new layouts c created for each chosen pair, the mutation probability m, and the number of generations G. Fig. 3 shows a visual representation of the algorithm.

2.5. Simulating lighting energy consumption based on building layouts

There are several viable simulation models for evaluating the expected energy consumption of different building layouts. These

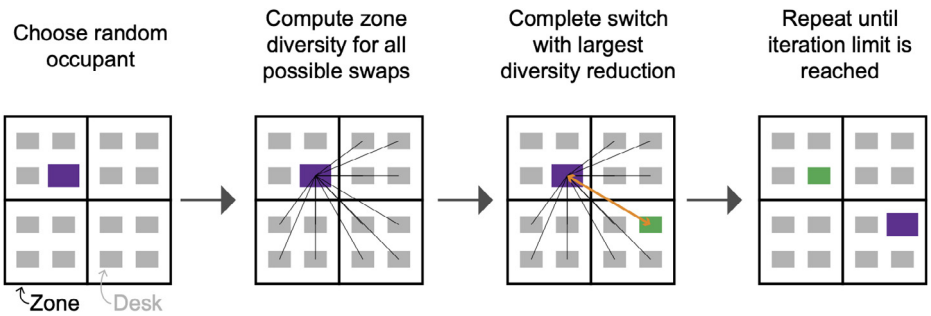

Fig. 2. Occupant clustering algorithm.

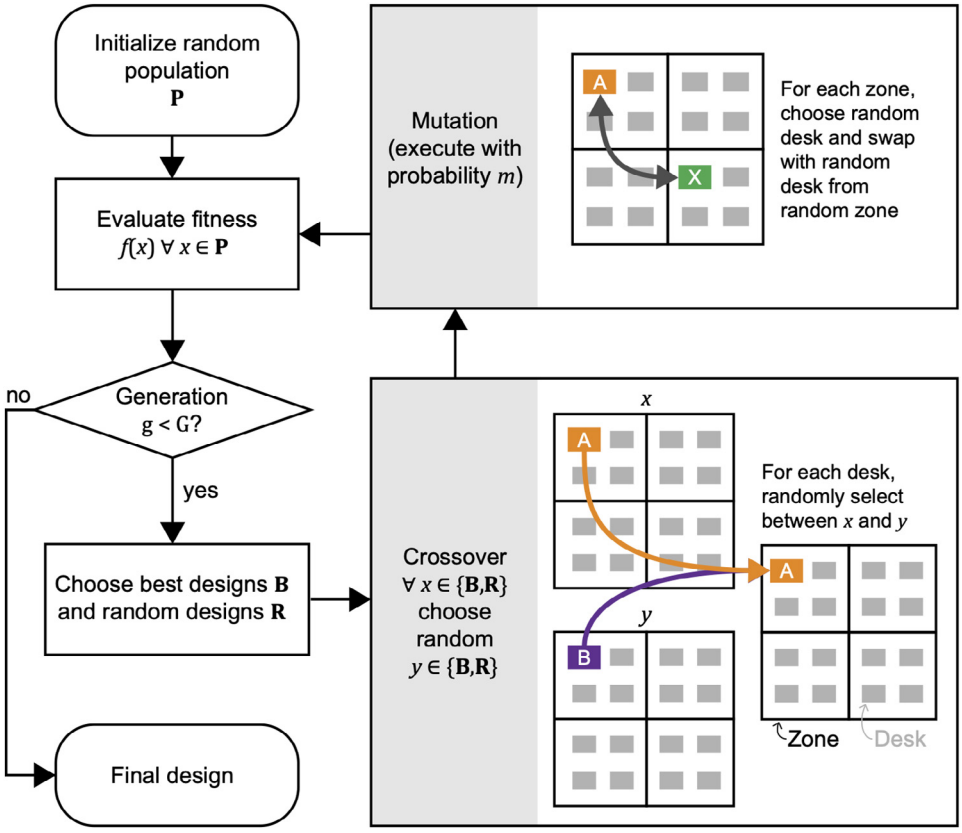

Fig. 3. Genetic algorithm adapted for building layout optimization.

models fall into two main categories: physics-based thermodynamic models (e.g., EnergyPlus [32]) and data-driven "surrogate" models [33]. Thermodynamic models have been shown to be particularly helpful when evaluating energy consumption from heating and cooling systems, though recently mixed models and data-driven surrogate models have become more prevalent. However, because our analysis focused on lighting systems with direct control through occupancy sensors (as explained below), the thermodynamic models are more complex than necessary for this task. Additionally, the time required for their analysis is prohibitive for running our genetic algorithm optimization. Therefore, we sought to develop a data-driven surrogate model that utilizes machine learning to predict lighting energy consumption based on occupant schedules as well as standard time series features (e.g., time of day, day of week, etc.). We chose to test multiple linear regression (MLR), support vector regression (SVR), random forests (RF), and artificial neural networks (ANN) to determine the most robust surrogate model for our purpose.

The 7 specific features we used for prediction of energy consumption were as follows:

- s_1, s_2, s_3: number of occupants in each of the three energy states as defined above in Section 2.1.
- Hour of day (0–23)
- Day of week (0–6)
- Weekday/weekend indicator (0 or 1)
- Zone number (0–number of zones)

Both the day of week and weekday/weekend indicator were included as features. Since these features are correlated with one another, the MLR model may be impacted by multicollinearity. We note that for the MLR model, this collinearity does not impact the power of prediction, but may make any hypothesis testing on the importance of features less reliable. The zone number was included to enable the model to adapt to zone-specific operational tendencies. For example, many lighting systems include daylight sensors, whereby the artificial lighting levels are lowered if enough daylight is present. This modulation would be expected to vary throughout the day for each zone, depending on orientation and other factors.

We implemented each potential surrogate model, described briefly here, using the scikit-learn package in Python [34].

- *Multiple linear regression.* The simplest model, MLR seeks to predict energy consumption ($\hat{Y}$) as a linear combination of the model features (X): $\hat{Y} = \beta X + \varepsilon$, where β is a vector of parameters and ε is the error term. The fitting of the parameters involves minimization of the error term.
- *Support vector regression.* The model produced by SVR relies on a small subset of the training data known as support vectors. Errors within the bounds created by the support vectors (within margin ε) are ignored. Fitting an SVR model involves the following optimization:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\zeta_i - \zeta_i^*\right)$$

subject to the following constraints:

$$y_i - w^T\phi(x) - b \le \varepsilon - \zeta_i$$

$$w^T\phi(x) + b - y_i \le \varepsilon - \zeta_i^*$$

$$\zeta_i^{(*)} \ge 0$$

where w is the set of feature weights ζ_i and ζ_i^* are the residuals beyond ε and φ is a kernel function that is often nonlinear such as the Gaussian radial basis function. The SVR model has been previously applied to building energy prediction tasks with success [35].

- *Random forests.* The RF regression model is an extension of the decision tree model in regression form, in which the overall model aggregates (generally through averaging) the result from many independently fit trees. Each tree constitutes a series of decisions on the features (e.g., time of day is less than or greater than 6 am), and once the full series of decisions are made, a final value is chosen. RF models have successfully been applied to energy prediction in various settings [36,37].
- *Artificial neural network.* An ANN is an interconnected group of nodes, in which each node produces a signal according to the data it receives. ANN architecture generally involves an input layer (which receives the features), one or more hidden layers, and an output layer (which produces a prediction). These networks are fit using backpropagation. They have been widely used for energy prediction tasks [37,38].

Each of the four models can be evaluated by splitting the data into a training set and a test set, and then performing 5-fold cross validation on the training set for model development. In this work, the initial split into training and test sets preserved the time series attribute of the data: the first 80% of the data, in terms of time, were used as the training set and the second 20% were used for testing. We preserved the time series attribute of the data in order to test the ability of the models to predict future events and also to enable time series visualization of the performance of each model. Once each of the four models was evaluated, we tuned the hyperparameters of the high-performing models. This hyperparameter tuning was again performed using 5-fold cross validation on the training set.

For the MLR, ANN, and SVR (non-tree-based) models, one-hot encoding was used for the day of week and zone number features, bringing the total number of features up to 23 features. Additionally, for these models, the hour of day feature was decomposed using sine and cosine transformations, to preserve the cyclical nature of the hour features (i.e., hour 23 is close to 0). Furthermore, the "state count" features (number of occupants in each energy state) were transformed using a sigmoid function. The intuition behind this transformation is that there are diminishing effects of having more than one occupant present in the zone (as shown in Fig. 4). This phenomenon occurs due to the fact that the lights only need to sense one occupant in order for all lights within a zone to turn on. Finally, the features were scaled to range between 0 and 1. Each of these steps were found to enhance the performance of the non-tree-based predictive models. These additional feature scaling steps were not used for the tree-based RF regressor, as the decisions are invariant to the scaling.

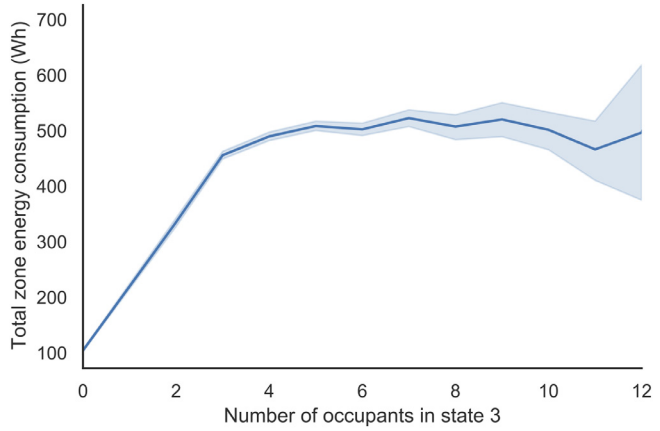

Fig. 4. Total zone energy consumption vs. number of occupants in state 3. The plateau indicates that once a small number of occupants are in the zone, the addition of more occupants does not further increase energy consumption.

2.6. Empirical data

We installed plug load energy sensors at each desk on a single floor or a large commercial office building in Redwood City, CA. The floor comprises 164 desks, of which 151 were occupied. The data collection period started August 1, 2019 and ended February 29, 2020. During the data collection period, the building operated under flexible standard work hours (i.e., most occupant activity occurred between the hours of 7 am and 7 pm, but with flexibility in terms of when occupants chose to use the building). Each occupant was assigned to their own workstation, which did not change during the study period. Occupants generally conducted standard office work, consisting of individual work at workstations as well as meetings with other occupants within and outside the study building. The sensors are Zooz SmartPlugs that communicate to a Samsung SmartThings hub through Z-Wave technology. The plug load sensors reported power consumption values any time the power consumption varied by more than 0.1 W. Consistent with previous work [25], we aggregated the power consumption to 15-minute intervals, as modeling space use at this time frequency has been shown to limit noise while providing useful information in terms of building operation. As described above in Section 2.1, we mapped the raw plug load data to energy states describing occupant schedules.

The office building is equipped with a lighting system that operates based on occupancy sensors, daylighting sensors, and schedules. The lighting zones are controlled by occupancy sensors across the building floor. If any lighting fixture within a zone senses motion over the past 20 min (10 min on weekends), all fixtures within that zone turn on. Eleven of the lighting zones service all workspaces (the others service other small shared spaces such as meeting rooms, corridors, etc.), and all lighting zones that service workstations use the same fixtures. We restricted our analysis to the 11 zones that service workspaces, as we were interested in characterizing energy consumption in places where individual schedules can be modeled according to our data. The lighting energy data is available at 1-hour intervals for the full duration of the study.

Due to persistent sensor outages at the beginning of data collection, our analysis began with data on October 1, 2019. In addition, due to a temporary shutdown of organizational activities over the New Year, we discarded data from December 16, 2019 to January 4, 2020. We therefore restricted our analysis to 132 full days of sensor and lighting energy data.

3. Results

3.1. Increased zone diversity correlates with increased energy consumption

For each day over the data collection period, and for each of the 11 lighting zones in the building, we computed the zone diversity, as discussed above in Section 2.2. We also computed the average energy consumption across lighting fixtures within a zone. We then completed a regression analysis for the relationship between zone diversity and energy consumption, as shown in Table 1. We

Table 1
Results of regression between lighting energy consumption and zone diversity across the 11 zones in the test building.

Zone	Number of Workstations	Coefficient (Standard Deviation)	*t*-statistic	*p*-value	R^2
0	16	422.5(78.29)	5.396	<0.001	0.165
1	16	492.4(52.64)	9.354	<0.001	0.354
2	10	828.2(106.7)	7.762	<0.001	0.288
3	10	532.4(80.00)	6.655	<0.001	0.288
4	20	527.9(35.33)	14.944	<0.001	0.627
5	6	749.6(126.3)	5.934	<0.001	0.253
6	16	514.3(40.78)	12.611	<0.001	0.288
7	16	989.2(51.83)	19.087	<0.001	0.666
8	22	177.6(39.70)	4.472	<0.001	0.117
9	22	486.9(20.94)	23.256	<0.001	0.757
10	10	97.1(60.90)	4.878	<0.001	0.157

found that for each zone, there was a positive relationship between zone diversity and energy consumption. We computed the *t*-statistic for the regression coefficient, and we found that the *p*-value for the *t*-statistic was significant at the 0.001 level for all zones. While the positive relationship is clear, both the slope of the relationship and strength of that relationship in terms of the R^2 value varied across zones. We note that each zone has a different size in terms of number of workstations within the zone, but we found that the differences in relationship strength do not seem to depend on this zone size. In Fig. 5a, we show the data along with the regression lines and 95% confidence intervals for each zone, and we show in Fig. 5b and c the individual regressions for the zones with the largest and smallest regression coefficients, respectively: Zone 7 and Zone 10. We note that this result serves as validation of the hypothesized relationship between occupant schedules and energy consumption. In addition, it serves as motivation for optimizing building layouts in order to reduce this diversity and therefore save energy.

3.2. Lighting energy simulation is driven by occupant schedule data

We tested the four models described above in Section 2.5 by performing 5-fold cross validation on the training set. We found that, before tuning of hyperparameters, the ANN performed the best in terms of mean squared error (MSE) and explained variance (R^2), and the RF model performed the best in terms of mean absolute error (MAE) (Table 2). Because the squared error terms, MSE and R^2, exaggerate the importance of larger errors, we can infer that the RF model occasionally produced larger errors, while the ANN model produced a higher baseline of error values. While the ANN model took the longest time to train, it required less computation time to produce a prediction than the RF model. Based on these results, we chose the ANN and RF models for additional hyperparameter tuning.

Our hyperparameter tuning for both models also involved 5-fold cross-validation on the training set. For the ANN model, we tuned the hidden layer sizes, activation function, solver, and learning rate. We performed a grid search on these hyperparameters and found the best parameters to be a single hidden layer of size 100, the tanh activation function, the Adam solver, and a learning rate of 0.01.

For the RF model, we tuned the number of trees, minimum number of samples to produce a split in the tree, the minimum number of samples per leaf, the maximum depth of the tree, and whether the bootstrap methodology was used in model training. We found the best parameters to be 200 trees, minimum split size of 50, minimum samples per leaf of 2, maximum depth of 300, and use of bootstrap in model training.

We found that our tuned RF model outperformed our tuned ANN, with an MAE of 6.27, MSE of 87.1, and R^2 of 0.715 as estimated through cross validation. We also note that while hyperparameter tuning did improve the performance of each model, in

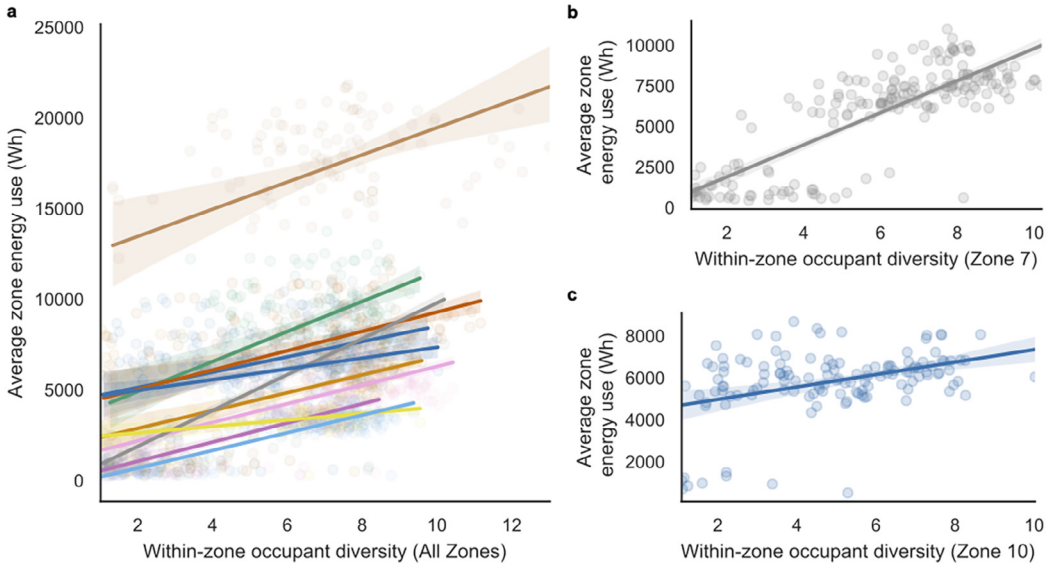

Fig. 5. Relationship between zone diversity metric and energy consumption along with regression fits and confidence intervals for (a) all zones—with colors representing different zones, (b) zone with the largest regression coefficient (Zone 7), and (c) zone with the smallest regression coefficient (Zone 10).

Table 2
Model results on 5-fold cross-validation.

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Explained Variance (R^2)	Time for Training (s)	Time for Prediction (s)
Multiple Linear Regression	9.55	141	0.534	0.0311	0.00198
Support Vector Regression	7.13	118	0.614	30.9	4.38
Random Forest Regression	6.11	98.2	0.678	2.82	0.0983
Artificial Neural Network	6.29	88.7	0.710	54.8	0.0105

both cases the improvement was very small. This high performance "out of the box" could enable wide adoption of similar surrogate models for the purposes of simulating energy consumption based on occupancy data.

Final results for the ANN and RF model, on both cross-validation and the test set, are shown in Table 3. Because the ultimate goal of our model is to be able to predict total energy consumption, as a result of features on the scale of 15 min, we also tested the performance of the ANN and RF models on a more aggregate level. We summed predicted and actual energy consumption to daily values and computed the R^2 metric on this aggregate level. As expected, both models benefited, in terms of their error rates, from this aggregation. However, the RF model benefited even further than the ANN. In Fig. 6, we show the predicted energy consumption (using the tuned RF model) vs. actual energy consumption data for the first seven days of data in the test set for zone 1. As this time-series plot shows, the large "jumps" in energy consumption (i.e., from low values to high values and vice versa) generally matched between the predicted and actual data, with these jumps likely driven by changes in occupancy. Much of the error seems to have resulted from small differences rather than large mischaracterizations by the model.

An additional strength of the RF model is its interpretability. Each split in a tree involves a decision on one of the features, which do not need to be scaled or one-hot encoded. As a result, the importance of each feature can be easily quantified and visualized. For the final RF model, we found the most important features to be time of day, and number of occupants in state 3 (Fig. 7). This importance of the occupant feature underscores the notion that the lighting system operation is driven by occupant behavior, and therefore there exists opportunity to save energy by adapting the layout of the building to the behavior of occupants.

We note that the time for prediction using the RF model was roughly one order of magnitude above the time for prediction using the ANN model. In our case, these times were sufficiently small such that the difference between using them was inconsequential to our analysis. However, in situations with particularly high computing costs, the ANN model, while slightly less accurate, could be used to improve computational costs. With that said, the RF model will always maintain a much higher level of interpretability.

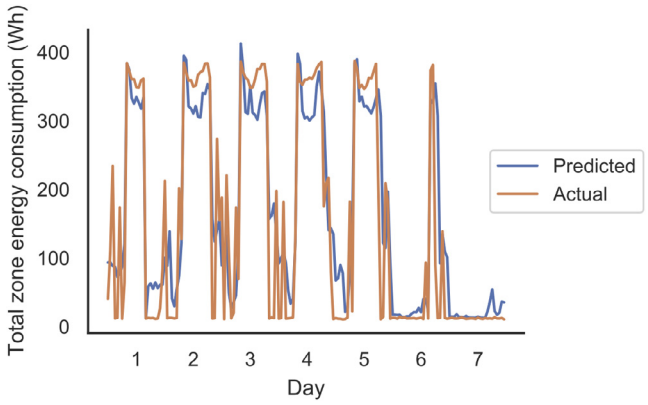

Fig. 6. Example predicted (using tuned RF model) vs. actual energy consumption data for the first seven days of data in the test set for zone 1.

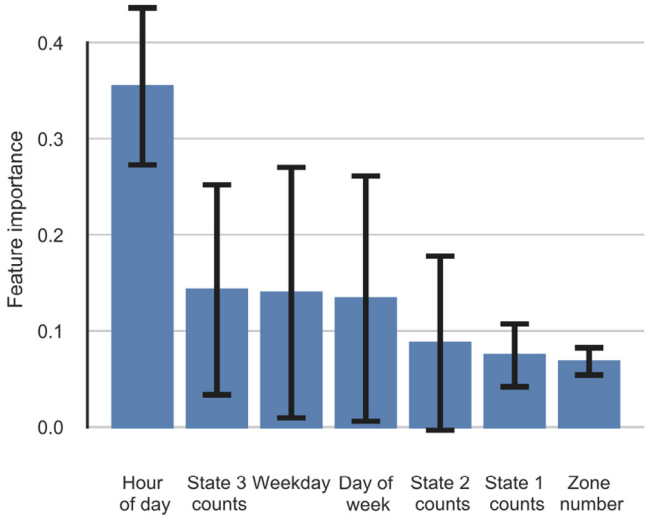

Fig. 7. Feature importance for final tuned random forest regression model.

Table 3
Model results after hyperparameter tuning on both 5-fold cross-validation and final test set.

Model	Errors on CV			Errors on Test Set	
	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Explained Variance (R^2)	Explained Variance (R^2) Hourly	Explained Variance (R^2) Daily
Tuned Random Forest Regression	6.27	87.1	0.715	0.740	0.834
Tuned Artificial Neural Network	6.28	88.5	0.710	0.734	0.817

3.3. Optimizing layouts reduces energy consumption according to simulations

In this section, we show the results of our clustering-based and genetic algorithm optimization methods on two examples: a synthetic example that demonstrates the underlying mechanics of the optimization routines and the surrogate simulation model, as well as the empirical data collected from the office building in Redwood City.

3.3.1. Synthetic example

Here, we introduce a simple example of an office building based on the hypothetical example from the introduction, though updated to be more representative of common office behavior. Our purpose in showing this synthetic example is to demonstrate the mechanics of the optimization methods in creating new building layouts based on occupant activities. In this synthetic example, there are four different "archetypes" of occupant schedules. Each archetype defines the following behaviors:

- **Arrival.** Time when occupant arrives at the building, transitioning from low energy state to a medium or high energy state.
- **Lunch.** Time and length of occupant's lunch, in which the occupant transitions to a low energy state for the duration of lunch.
- **Meeting(s).** Time(s) and length(s) of the occupant's meeting(s), in which the occupant transitions to a low energy state for the duration of the meeting(s).
- **Departure.** Time when occupant leaves the building, transitioning to a low energy state.

When the occupant is in a normal working state, between arrival and departure but not during lunch or a meeting, the occupant randomly transitions between high and medium energy states. This behavior models normal occupant behavior of taking short breaks throughout the workday. The four specific occupant archetypes used in this synthetic example are shown in Table 4.

The hypothetical building includes four rooms, with nine desks per room. Each occupant follows one of the four archetypes and there are exactly 9 of each archetype. We simulated energy consumption for one day, leveraging the random forest model discussed above in Section 2.5. For simplicity, we assigned each room to be the same zone number (0), which in the empirical example, has roughly the same size of 9 occupants. We also assumed the day was a Monday. We tested two different layouts to show the effects of layout on energy consumption: a random layout, and a layout in which all 9 occupants of each archetype share a room. We found, leveraging the lighting energy surrogate simulation model trained on real-world data, that the archetype-

Table 4
Occupant archetypes used in the synthetic example.

Archetype	Arrival	Lunch	Meetings	Departure
1	9 am	12 pm (1 h)	3 pm (1 h)	5 pm
2	9 am	None	None	4 pm
3	11 am	3 pm (1 h)	3 pm (1 h)	7 pm
4	7 am	11 pm (1 h)	1 pm (2 h)	5 pm

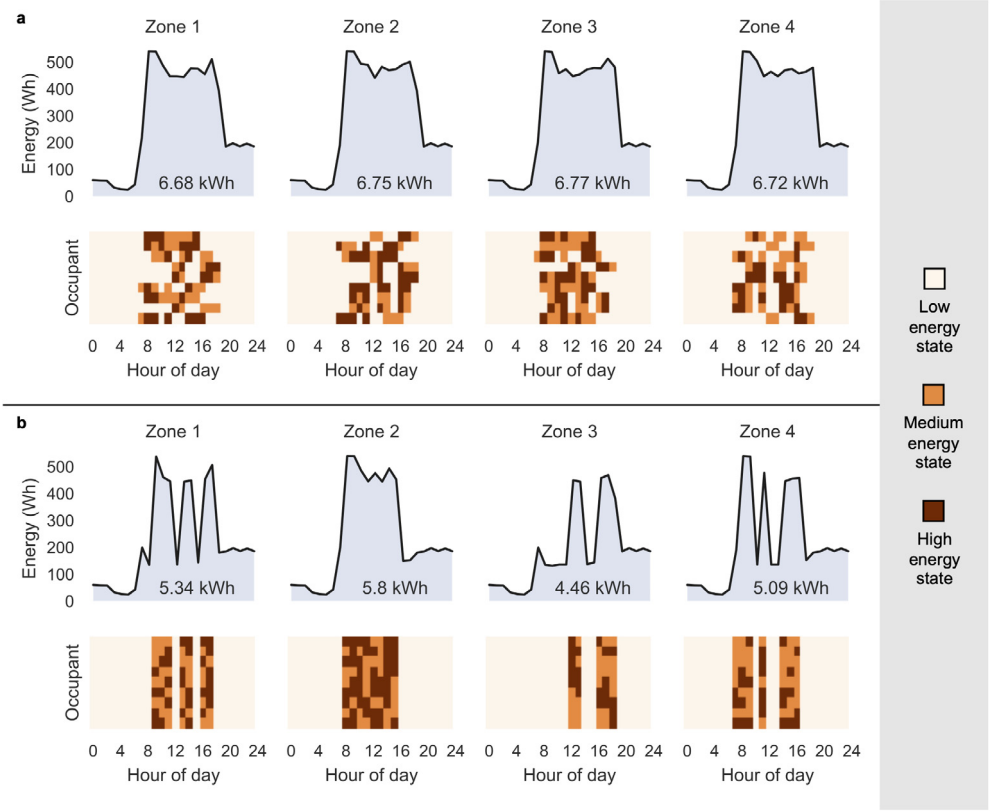

Fig. 8. Synthetic example occupant schedules and simulated energy consumption for (a) random layout and (b) known optimal layout.

based layout created a situation in which the lighting system was able to respond to occupants' use of the space (Fig. 8). The result was significant energy savings of 22%.

We know by inspection that the optimal layout is the layout in which the building is arranged by archetype. This *a priori* knowledge of the system enables evaluation of the optimization routines introduced in Sections 2.4 and 2.5. When we applied the clustering-based optimizer to this example, we found that the optimizer was able to arrive at the known optimal layout very quickly (Fig. 9). The genetic algorithm, on the other hand, quickly found a near-optimal solution, but was unable to reach the fully optimal layout. This behavior of genetic algorithms—in which they come close to optimality but have final convergence issues—is well documented [39]. However, this synthetic example is relatively simple. In a real-world office building, with much more variation in occupant activities both in time and space, we would expect the genetic algorithm to provide a sufficient solution, as previous work has demonstrated that such methods can avoid local minima with sufficient variation in input data [39].

3.3.2. Empirical case study

The synthetic example offers insight into the mechanics of the optimizers developed in this work, but the overly simplistic nature of that example limits claims that may be made about the extensibility of our framework. We therefore leveraged the optimization algorithms and surrogate simulation model to optimize the layout of the Redwood City office building introduced in Section 2.6. We simulated energy consumption using the full dataset of 132 days of occupant behavior. We started by estimating the energy consumption of the lighting system from 100 random building layouts as well as the existing building layout. We then applied the clustering-based optimization routine to the data with varying dimensionality (3, 5, 10, 30, 100, 151, and full dimensionality without reduction), as well as the genetic algorithm. We ran each optimizer 100 times to create 100 layouts and ultimately a distribution of estimated energy consumption (Fig. 10). Overall, we found that the existing layout performs slightly better than the random layouts, but that further improvements on energy consumption could be realized by optimizing the layouts for energy efficiency. Additionally, increasing the number of dimensions used in the clustering optimizer improved performance.

We also computed the zone diversity in occupant schedules as introduced in Section 2.2. Here, the zone diversity metric was computed for the entire 132 days, as opposed to daily in Section 2.2. In general, we found that the zone diversity metric followed the predicted energy consumption (Fig. 11). However, we note that while the layouts produced through the genetic algorithm were among the best, the zone diversity metrics for the genetic algorithm layouts were substantially higher than for the clustering-based layouts.

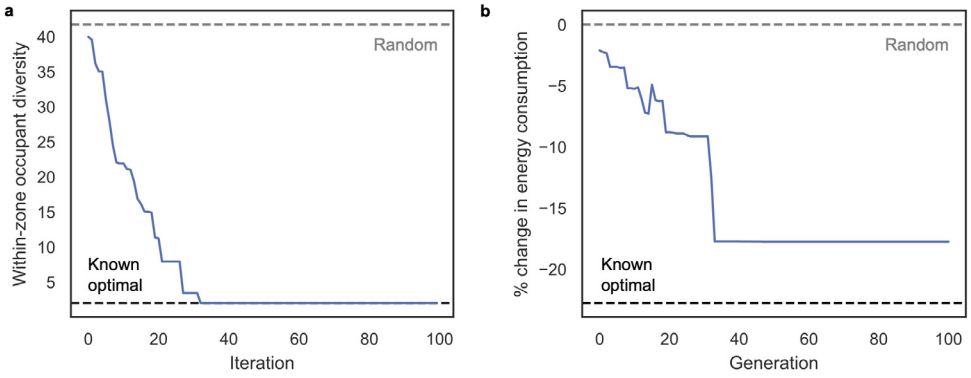

Fig. 9. Optimization convergence on synthetic example, in relation to a random layout and the known optimal layout for (a) clustering-based optimization and (b) genetic algorithm. The same optimal layout produces the zone diversity value and energy values shown in the figure.

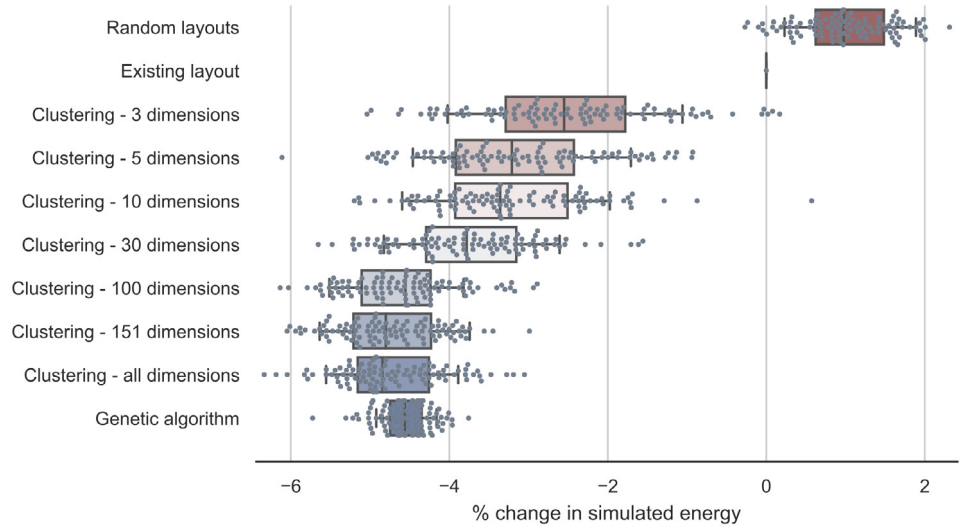

Fig. 10. Simulated energy consumption (expressed as % change from the existing layout) for random and optimized building layouts.

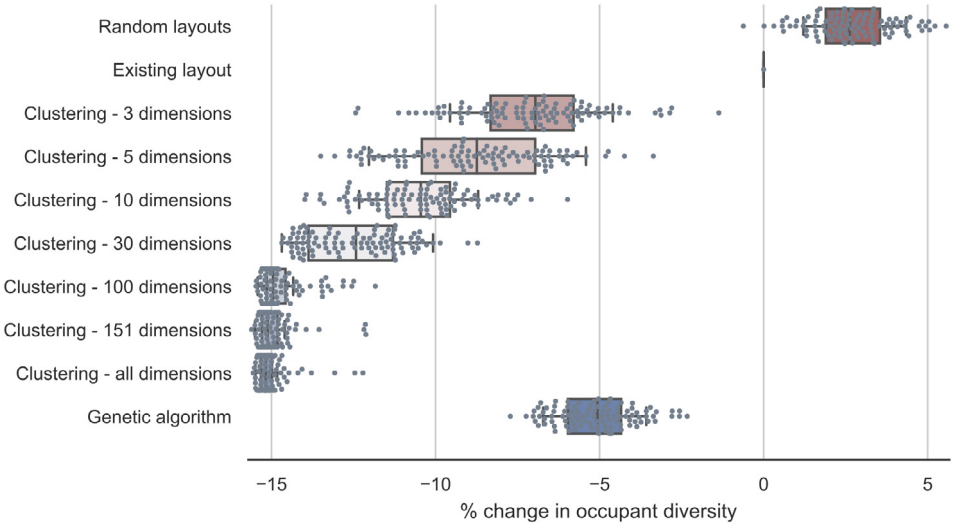

Fig. 11. Change in zone diversity metric (expressed as % change from existing layout) for random and optimized building layouts.

4. Discussion

Our results indicate the possibility for significant energy savings of the lighting system through redistribution of occupants in space. This energy savings is possible by leveraging the occupancy-sensing feature of the lighting system. We note that while the heating, cooling, and ventilation system was not modeled in our case study due to data limitations, we would expect to find similar results as long as the thermal system is able to reduce service during periods without occupancy. The analysis of expected effects on these additional building systems is a rich area for further work.

While optimization creates reductions in energy consumption compared to the existing layout, it is interesting to note that the existing layout has a smaller simulated energy consumption than ~90% of the random layouts. This is not altogether surprising, as it indicates that behavior in the existing building is more aligned than if each occupant behaved independently. This result may suggest that people tend to adapt to the behavior of those around them, as has been documented in previous social science research [40]. This question of how people adapt their behavior to their surroundings has potentially large implications for interpretations of our results. Underlying our analysis is the assumption that individual behavior will not change when occupant layouts are changed. We note that this assumption is a limitation of our model, and that behavior can be expected to change in some way in response to new desk assignment changes. The direction of the impact of these behavioral changes on energy consumption is unclear. However, our finding that the existing layout is more efficient than a random layout suggests that people adapt their behavior to match those around them. It is therefore possible that when we create layouts in order to reduce within-zone diversity in behavior, occupants may naturally choose to adapt their behavior and therefore further reduce within-zone diversity. On the other hand, it may be possible that occupants may choose to alter their behavior in other ways, either by attempting to maintain social behaviors that precipitated out of their previous setting, or perhaps by choosing to separate themselves from the others in their new surroundings. We note that the evidence from this paper seems to suggest that individuals, at least to some extent, tend to assimilate their behavior to those around them.

A key finding from this work is that the reduction in energy consumption from both the clustering-based optimization (when enough dimensions are used) and the genetic algorithm is roughly the same (Fig. 10). This is particularly interesting because the clustering-based optimization procedure does not include any information about the surrogate model used for predicting energy consumption, whereas the genetic algorithm uses that model as feedback during its execution. We argue that one can reasonably view the genetic algorithm as closer to a "best-case scenario" for predicted energy consumption, given its explicit use of simulation as a feedback mechanism during optimization. Our finding that clustering (once enough dimensions are used to represent the data) performs just as well demonstrates the strength of the naïve clustering-based approach. When only a small number of dimensions are used for clustering, the optimization does result in energy reduction, though the effect is smaller. It is therefore important to ensure that enough dimensions are used for representation of the occupancy data.

The effect of clustering on the within-zone occupant diversity metric is clear: substantial diversity can be reduced by clustering occupants according to their schedules. The genetic algorithm also reduces this zone diversity, but to a significantly smaller degree (Fig. 11). This finding suggests that there are other factors beyond zone diversity that are addressed during execution of the genetic algorithm. We unfortunately cannot interpret what these factors are, but they could involve unique aspects of the lighting system's operation. It is also possible that the genetic algorithm is optimizing for uncertainty in the random forest surrogate simulation model, which could be a limitation of that approach.

We found that the clustering-based optimization performs about as well as the genetic algorithm. The genetic algorithm is only executable when a simulation engine is available, which makes the clustering-based approach more practical in situations when only occupancy data are available. We note, however, that if a simulation model is available, the genetic algorithm can be seeded with the layouts obtained through clustering. In our case study, the clustering-based layouts form a distribution with regard to simulated energy consumption (Fig. 10). In a preliminary analysis, we found that seeding the genetic algorithm with 50 clustering-based layouts and 50 random layouts created new layouts that performed as well as the best layouts from the clustering (the layouts furthest to the left on the distribution in Fig. 10). Therefore, this ensemble approach may be useful in reducing uncertainty around the expected outcomes from either approach.

As discussed in Section 1.3, optimization of building design and system control can create significant energy saving opportunities

in buildings. For example, Krioukov et al. [15] found that occupant-driven control of lighting systems can lead to 50–70% energy savings. These strategies are critical for making use of controllable building systems and achieving energy efficiency. In some ways, these strategies can be seen as the "low hanging fruit" of optimal building performance (e.g., turning the lights off when no occupants are present). The research we present is this paper takes these strategies one step further: our methodology is not just about optimizing building system operation using information about occupant dynamics, but also optimizing characteristics of those occupants—the layout of workstations—to take full advantage of building systems. The approach we introduce can be considered in tandem with the more traditional approaches of controlling energy-intensive building systems, and it has the added benefit of not requiring expensive upgrades to building automation systems in order to achieve additional energy savings. Furthermore, the approach we introduce can also be integrated with other characteristics of building design, such as thermal comfort or organizational success, both of which can be tied to building layout [41,42].

The ultimate goal of the design optimization in this paper is to reduce energy consumption, a goal that is important for reaching our sustainable energy goals and reducing costs for organizations. However, it is essential to note that there are other goals for the success of the building and organization that should be considered as well. Chief among these goals is productivity—perhaps the ultimate purpose of commercial office buildings—which is difficult to define and varies among different organization. Numerous factors have been shown to influence productivity, including occupant thermal comfort [43], organizational cohesion, and others. The benefit of the occupant-driven optimization-based design approach we introduce here is that there is a natural extension to include other objectives (such as those discussed above) through a multi-objective optimization. We specifically note that to address organizational cohesion and collaboration, the approach of space syntax analysis [42,44] may yield viable optimization objectives. Other graph theoretical approaches that are designed to leverage organizational needs [23] could also complement the approach we introduce here. The integration of occupant-centric design for energy with design for organizational outcomes will be a rich area of future research investigations.

In practice, any theoretical improvements—whether for energy performance or organizational performance—that depend on making changes to buildings' physical layouts or seating arrangements may elicit hesitation on the part of the building manager. We note, however, that such spatial reorganizations are not uncommon in many modern workplaces, and some workplaces may be trending toward changing seating arrangements more frequently [45,46]. As offices begin to realize the potential collaborative benefits of making such changes, the framework presented in this paper can be a means to integrate the energy perspective into organizational design. We believe that one of the strengths of this research is that it sheds light on the power of investigating characteristics of building layouts—aspects of building design and management that may be too often overlooked. In other words, once building and organization managers better understand how layouts impact energy consumption and the workplace experience, they may be more willing to make changes. If the occupant arrangement is viewed as an energy-saving investment, it may be seen as a less-expensive alternative to other investments.

One limitation of this work is that not all buildings use fixed workstations for their employees. As discussed in Section 1 above, some offices use activity-based workplace (ABW) design that allows workers to choose different locations for different tasks. As office dynamics change in response to workplace norms, post-Covid-19 worker and management preferences, and other factors, different approaches to office spatial management will work best for different offices. We note, however, that fixed workstation seating arrangements continue to be most common in existing buildings. Furthermore, as more flexible spatial strategies grow in prevalence, the specific optimization routines can evolve in tandem. For example, a more general term describing overall "space-use diversity" rather than "individual occupant zone diversity" could be used to evaluate the space-use patterns of a flexible office, and this information could be used to adjust aspects of office design specific to flexible offices, such as the amount of each type of space and its location in the building.

5. Conclusion

In this paper, we explored the relationship between occupant behavioral dynamics and energy consumption from energy-intensive building systems. We introduced (1) a zone diversity metric adapted from the literature for comparison with empirical building energy data, (2) a clustering-based building layout optimization methodology made possible through dimensionality reduction, (3) a novel genetic algorithm for building layout optimization, and (4) a data-driven surrogate simulation engine for predicting lighting energy consumption from occupancy data. In a case study, we found a significant relationship between building occupant zone diversity and actual lighting energy consumption. We also found that our layout optimization methods can be expected to reduce lighting energy consumption by about 5% from the existing layout and 6% from a random layout. Overall, we show that reconsidering the design of layouts in existing buildings has significant potential for realizing energy savings. Additionally, the approach of changing layouts to achieve energy efficiency also enables the simultaneous consideration of other factors influence by building layout, including organizational performance and thermal comfort. These methods, when integrated with the many objectives that drive building management, will be critical to ensuring dynamic and lasting energy savings in buildings.

CRediT authorship contribution statement

Andrew Sonta: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Visualization. **Thomas R. Dougherty:** Methodology, Software, Writing - original draft. **Rishee K. Jain:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition.

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.

Acknowledgements

This work has been supported by a Stanford Graduate Fellowship, the Center for Integrated Facility Engineering, a Terman Faculty Fellowship, and the United States National Science Foundation (NSF) under Grant Nos. 1836995 and 1941695. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

References

[1] Z. Yang, A. Ghahramani, B. Becerik-Gerber, Building occupancy diversity and HVAC (heating, ventilation, and air conditioning) system energy efficiency, Energy 109 (2016) 641–649, https://doi.org/10.1016/j.energy.2016.04.099.

[2] O. Lucon, D. Ürge-Vorsatz, A. Zain Ahmed, H. Akbari, P. Bertoldi, L.F. Cabeza, N. Eyre, A. Gadgil, L.D.D. Harvey, Y. Jiang, E. Liphoto, S. Mirasgedis, S. Murakami, J. Parikh, C. Pyke, M.V. Vilariño, Buildings (2014), https://doi.org/10.2753/JES1097-203X330403.
[3] J.P. Basbagill, F.L. Flager, M. Lepech, A multi-objective feedback approach for evaluating sequential conceptual building design decisions, Autom. Constr. 45 (2014) 136–150, https://doi.org/10.1016/j.autcon.2014.04.015.
[4] B. Roisin, M. Bodart, A. Deneyer, P. D'Herdt, Lighting energy savings in offices using different control systems and their real consumption, Energy Build. 40 (4) (2008) 514–523, https://doi.org/10.1016/j.enbuild.2007.04.006.
[5] S. Petersen, S. Svendsen, Method and simulation program informed decisions in the early stages of building design, Energy Build. 42 (7) (2010) 1113–1119, https://doi.org/10.1016/j.enbuild.2010.02.002.
[6] H. Tanikawa, S. Hashimoto, Urban stock over time: spatial material stock analysis using 4d-GIS, Build. Res. Inf. 37 (5-6) (2009) 483–502, https://doi.org/10.1080/09613210903169394.
[7] X. Feng, D.a. Yan, T. Hong, Simulation of occupancy in buildings, Energy Build. 87 (2015) 348–359, https://doi.org/10.1016/j.enbuild.2014.11.067.
[8] A. Roetzel, A. Tsangrassoulis, U. Dietrich, Impact of building design and occupancy on office comfort and energy performance in different climates, Build. Environ. 71 (2014) 165–175, https://doi.org/10.1016/j.buildenv.2013.10.001.
[9] W. Parys, D. Saelens, H. Hens, Impact of occupant behaviour on lighting energy use, Build. Simul. (2009) 1143–1150.
[10] L.K. Norford, R.H. Socolow, E.S. Hsieh, G.V. Spadaro, Two-to-one discrepancy between measured and predicted performance of a 'low-energy' office building: insights from a reconciliation based on the DOE-2 model, Energy Build. 21 (2) (1994) 121–131, https://doi.org/10.1016/0378-7788(94)90005-1.
[11] M. Jia, R.S. Srinivasan, A.A. Raheem, From occupancy to occupant behavior: an analytical survey of data acquisition technologies, modeling methodologies and simulation coupling mechanisms for building energy efficiency, Renewable Sustainable Energy Rev. 68 (2017) 525–540, https://doi.org/10.1016/j.rser.2016.10.011.
[12] J.Y. Park, M.M. Ouf, B. Gunay, Y. Peng, W. O'Brien, M.B. Kjærgaard, Z. Nagy, A critical review of field implementations of occupant-centric building controls, Build. Environ. 165 (2019) 106351, https://doi.org/10.1016/j.buildenv.2019.106351.
[13] W. Yu, B. Li, H. Jia, M. Zhang, D. Wang, Application of multi-objective genetic algorithm to optimize energy efficiency and thermal comfort in building design, Energy Build. 88 (2015) 135–143, https://doi.org/10.1016/j.enbuild.2014.11.063.
[14] P. Hoes, M. Trcka, J.L.M. Hensen, B.H. Bonnema, Optimizing building designs using a robustness indicator with respect to user behavior, in: Proc. Build. Simul. 2011 12th Conf. Int. Build. Perform. Simul. Assoc., 2011, pp. 1710–1717 (accessed May 22, 2020).
[15] A. Krioukov, S. Dawson-Haggerty, L. Lee, O. Rehmane, D. Culler, A living laboratory study in personalized automated lighting controls, BuildSys 11 (2011) 1–6, https://doi.org/10.1145/2434020.2434022.
[16] B. Balaji, J. Xu, A. Nwokafor, R. Gupta, Y. Agarwal, Sentinel: occupancy based HVAC actuation using existing WiFi infrastructure within commercial buildings, in: Proc. 11th ACM Conf. Embed. Networked Sens. Syst. - SenSys '13, ACM Press, Rome, Italy, 2013, pp. 1–14, https://doi.org/10.1145/2517351.2517370.
[17] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, T. Weng, Occupancy-driven energy management for smart building automation, in: Proc. 2nd ACM Work. Embed. Sens. Syst. Energy-Efficiency Build. - BuildSys '10, ACM Press, New York, New York, USA, 2010, p. 1, https://doi.org/10.1145/1878431.1878433.
[18] A. Ghahramani, K. Dutta, B. Becerik-Gerber, Energy trade off analysis of optimized daily temperature setpoints, J. Build. Eng. 19 (2018) 584–591, https://doi.org/10.1016/j.jobe.2018.06.012.
[19] C. Candido, L. Thomas, S. Haddad, F. Zhang, M. Mackey, W. Ye, Designing activity-based workspaces: satisfaction, productivity and physical activity, Build. Res. Inf. 47 (3) (2019) 275–289, https://doi.org/10.1080/09613218.2018.1476372.
[20] T. Sood, P. Janssen, C. Miller, Spacematch: using environmental preferences to match occupants to suitable activity-based workspaces, Front. Built Environ. 6 (2020) 113, https://doi.org/10.3389/fbuil.2020.00113.
[21] K. Song, S. Kim, M. Park, H.-S. Lee, Energy efficiency-based course timetabling for university buildings, Energy 139 (2017) 394–405, https://doi.org/10.1016/j.energy.2017.07.176.
[22] J.H. Jo, J.S. Gero, Space layout planning using an evolutionary approach, Artif. Intell. Eng. 12 (1998) 149–162 (accessed September 6, 2017) http://ac.els-cdn.com/S095418109700037X/1-s2.0-S095418109700037X-main.pdf?_tid=a2a8c5a8-931a-11e7-871b-00000aacb362&acdnat=1504713009_a3be26a0231c1d29633a827b88e4de22.
[23] J.I. Lather, T. Logan, K. Renner, J.I. Messner, Evaluating generated layouts in a healthcare departmental adjacency optimization problem, in: Comput. Civ. Eng., 2019, pp. 539–546, https://doi.org/10.1061/9780784482438.068.
[24] H.-Y. Lee, I.-T. Yang, Y.-C. Lin, Laying out the occupant flows in public buildings for operating efficiency, Build. Environ. 51 (2012) 231–242, https://doi.org/10.1016/j.buildenv.2011.11.005.
[25] A.J. Sonta, P.E. Simmons, R.K. Jain, Understanding building occupant activities at scale: an integrated knowledge-based and data-driven approach, Adv. Eng. Inf. 37 (2018) 1–13, https://doi.org/10.1016/j.aei.2018.04.009.
[26] J. Zhao, B. Lasternas, K.P. Lam, R. Yun, V. Loftness, Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining, Energy Build. 82 (2014) 341–355, https://doi.org/10.1016/j.enbuild.2014.07.033.
[27] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is "nearest neighbor" meaningful?, in: Int Conf. Database Theory, Springer, 1999, pp. 217–235.
[28] J. Leskovec, A. Rajaraman, J.D. Ullman, J. Leskovec, A. Rajaraman, J.D. Ullman, Dimensionality Reduction, in: Min. Massive Datasets, 2014. http://dx.doi.olrg/10.1017/cbo9781139924801.012.
[29] A.J. Sonta, R.K. Jain, Optimizing neighborhood-scale walkability, in: Comput. Civ. Eng. 2019, Atlanta, GA, 2020, pp. 454–461, https://doi.org/10.1061/9780784482438.058.
[30] D. Nagy, D. Lau, J. Locke, J. Stoddart, L. Villaggi, R. Wang, D. Zhao, D. Benjamin, Project discover: an application of generative design for architectural space planning, in: Proc. Symp. Simul. Archit. Urban Des., Society for Computer Simulation International, San Diego, CA, USA, 2017, https://doi.org/10.5555/3289787.3289794.
[31] S. RazaviAlavi, S. AbouRizk, Site layout and construction plan optimization using an integrated genetic algorithm simulation framework, J. Comput. Civ. Eng. 31 (4) (2017), https://doi.org/10.1061/(ASCE)CP.1943-5487.0000653 04017011.
[32] U.S. Department of Energy, EnergyPlus, (n.d.).
[33] A.N. Aijazi, L.R. Glicksman, Comparison of regression techniques for surrogate models of building energy performance, Proc. SimBuild 2016, 2016.
[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830 (accessed September 1, 2017) http://www.jmlr.org/papers/v12/pedregosa11a.html.
[35] R.K. Jain, K.M. Smith, P.J. Culligan, J.E. Taylor, Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy, Appl. Energy 123 (2014) 168–178, https://doi.org/10.1016/j.apenergy.2014.02.057.
[36] Z. Wang, Y. Wang, R. Zeng, R.S. Srinivasan, S. Ahrentzen, Random Forest based hourly building energy prediction, Energy Build. 171 (2018) 11–25, https://doi.org/10.1016/j.enbuild.2018.04.008.
[37] M.W. Ahmad, M. Mourshed, Y. Rezgui, Trees vs Neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption, Energy Build. 147 (2017) 77–89, https://doi.org/10.1016/j.enbuild.2017.04.038.
[38] B.B. Ekici, U.T. Aksoy, Prediction of building energy consumption by using artificial neural networks, Adv. n Eng. Softw. 40 (5) (2009) 356–362, https://doi.org/10.1016/j.advengsoft.2008.05.003.
[39] M.J. Kochenderfer, T.A. Wheeler, Algorithms for Optimization, first ed., MIT Press, 2018 (accessed November 26, 2018).
[40] T.L. Chartrand, J.A. Bargh, The chameleon effect: The perception-behavior link and social interaction, J. Pers. Soc. Psychol. 76 (1999) 893–910, https://doi.org/10.1037/0022-3514.76.6.893.
[41] S. Nagarathinam, A. Vasan, V. Sarangan, R. Jayaprakash, A. Sivasubramaniam, Good set-points make good neighbors - User seating and temperature control in uberized workspaces, in: Proc. 5th Conf. Syst. Built Environ., 2018, pp. 144–147, https://doi.org/10.1145/3276774.3276781.
[42] J. Peponis, S. Bafna, R. Bajaj, J. Bromberg, C. Congdon, M. Rashid, S. Warmels, Yan Zhang, C. Zimring, Designing space to support knowledge work, Environ. Behav. 39 (6) (2007) 815–840, https://doi.org/10.1177/0013916506297216.
[43] T. Keeling, D. Clements-Croome, R. Luck, P. Pointer, How the sensory experience of buildings can contribute to wellbeing and productivity, Proc. 7th Wind. Conf. Chang. Context Comf. an Unpredictable World, Network for Comfort and Energy Use in Buildings, Windsor, UK, 2012 (accessed August 30, 2017).
[44] S. Bafna, Space syntax: a brief introduction to its logic and analytical techniques, Environ. Behav. 35 (1) (2003) 17–29, https://doi.org/10.1177/0013916502238863.
[45] Why You Should Rotate Office Seating Assignments, Harv. Bus. Rev. (2018).
[46] S. Lee, Learning-by-moving: can reconfiguring spatial proximity between organizational members promote individual-level exploration?, Organ. Sci. 30 (3) (2019) 467–488, https://doi.org/10.1287/orsc.2019.1291.