Infomorphism: Integrating A Forecasting Model for High-Resolution Building Energy Demand Prediction

Fengqi Li$^1$, Kristen R. Kristen$^2$, Haolin Yang$^1$, Alexandros Tsamis$^1$

$^1$Rensselaer Polytechnic Institute, Troy, NY, United States
$^2$Carleton University, Ottawa, ON, Canada

Abstract

The idea of integrating renewable energy into urban design, through advanced building technologies and architectural design approaches, has recently started to see more attention. Accurately predicting energy consumption at a higher time resolution provides opportunities to broaden the scope of problems that can be addressed in planning frameworks. Increased granularity in time resolution, however, brings attendant problems such as increased data volatility, increased data volume, and non-linearity. In this work, these challenges are addressed through a machine learning model selection process, to best predict hourly building energy demand. The model is trained on five different building data sets. This work finds that, across all buildings tested, a Random Forest model provides the best predictive accuracy - up to a 78% improvement over comparable state-of-the-art models such as a dense deep learning network.

Highlights

- Model development to predict hourly building energy consumption
- Model development for both energy forecasting and prediction
- Random forest model provides best predictive accuracy
- Increased time resolution broadens applicability of Infomorphism planning framework

Introduction

A recently proposed planning framework, titled Infomorphism (Li et al. (2022)), embeds an energy demand forecasting model for building-level energy consumption while addressing system-level energy efficiency and urban planning issues related to renewable energy integration. The computational planning framework augments a generative planning process for building envelopes with a local energy-sharing network optimization model to explore potential planning policies associated with renewable energy equity. Taking renewable energy accessibility as a driver for optimizing planning envelopes, Infomorphism as an AI-based framework helps optimize energy efficiency for a city as a whole and balances energy exchange between areas of supply and areas of demand. Several case studies for Manhattan have been conducted to provide alternative planning environments for validating the effectiveness of the proposed framework and computational workflow.

The case studies show how a city can be developed as an energy network that ensures equitable access to renewable energy (heat and electricity) absorbed from the planning envelopes with minimum levelized energy costs. Establishing new policies and regulations according to equitable energy rights associated with renewable energy integration can collectively drive a city’s form, function, and infrastructure and discuss energy policies emerging from this research. It is anticipated that the development of Infomorphism will support the decision-making process related to architectural design, urban planning, energy infrastructure design, and renewable energy integration at both building and urban scales.

However, a key data constraint is the lack of publicly available datasets on building-level energy consumption, particularly at a high time resolution. Validation of the Infomorphism framework relied on the annual energy consumption of buildings, which was made available through the (NYC Office of Climate & Sustainability (2021)). In our previous paper (Li et al. (2022)), we presented an energy demand forecasting model that successfully predicted the annual consumption of a building. This model took into account various influential factors, including weather conditions that impacted solar availability and demand. Additionally, the model incorporated the building’s configuration and Floor Area Ratio (FAR), land use, and floor area, which were regulated by urban planning guidelines. Furthermore, the network pattern of the building, constrained by the street network, was also considered in the model. Through the incorporation of these variables, the developed forecasting model provided comprehensive insights into building energy consumption, facilitating informed decision-making in urban planning and energy management. In contrast, the energy consumption models developed in this work focus on finer time scales of energy demand data, enabling the analysis of a broader set of urban-energy questions which require a higher-fidelity of electric power flow, such as how the transition to electric vehicles and charging placement might affect the optimal energy network and urban form. Thus, a forecasting model based on higher resolution time data should be developed and integrated into the existing Infomorphism framework.

Proceedings of the 18th IBPSA Conference
Shanghai, China, Sept. 4-6, 2023
https://doi.org/10.26868/25222708.2023.1395
Building upon recent advances in databases that provide hourly energy demand at the building level (Hou et al. (2021), Sandberg et al. (2017), Kalhori et al. (2022)), this work develops a novel, hourly prediction of energy consumption for several different building types. The resulting machine learning model is then embedded into the larger Infomorphism framework, modeling urban renewable energy integration at higher resolution and higher fidelity.

Incorporating this new machine learning forecasting model into the overall urban design framework will enable the study, evaluation, and recommendation of urban energy policies at different resolutions. Of interest are policies related to energy equity and energy access, and how the formulation of such policies can enhance the overall urban form.

**Building Energy Data**

This study is based on building energy data collected continuously at different circuits within each building (Pecan Street Inc (2022)). The buildings modeled are located in New York state. Table 1 provides more context on the buildings, including construction year, square footage, and whether solar photovoltaic (PV) panels are installed on the building. If so, the total building energy consumption is less the solar PV output.

<table>
<thead>
<tr>
<th>Dataset No.</th>
<th>Year</th>
<th>Square Footage</th>
<th>Start Date</th>
<th>End Date</th>
<th>Solar PV Installed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1987</td>
<td>2,358</td>
<td>15-Feb-19</td>
<td>26-Feb-23</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>1940</td>
<td>1,368</td>
<td>19-Feb-23</td>
<td>26-Feb-23</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>1975</td>
<td>4,794</td>
<td>2-Apr-19</td>
<td>27-Feb-23</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>1940</td>
<td>1,883</td>
<td>2-Apr-19</td>
<td>7-Feb-23</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>1941</td>
<td>1,582</td>
<td>15-Feb-19</td>
<td>26-Feb-23</td>
<td>Yes</td>
</tr>
</tbody>
</table>

The energy data is available per minute, but in this study, the minute-level data is summed to the hourly level, which still retains the variance in energy consumption data. Thus, the data models the energy consumption of the entire building, per hour. An hourly time interval is required by the overall Infomorphism modeling framework, which hosts an embedded mixed-integer linear optimization problem (MILP) to develop local, building-to-building energy networks. To ensure the computational tractability of the MILP, which relies on the forecasts produced from the model proposed in this paper, the lowest time resolution possible is hourly.

Figure 1 shows the average energy consumption per hour, across the entire time series. Peak energy demand in each building occurs in the evening, between 17:00h to 21:00h. This figure also illustrates the variability in the data. Building 1, for example, sees large increase in energy consumption at 7:00h, and almost a correspondingly large decrease in energy consumption at 9:00h. Such frequent ramps in data values present a challenge in predictive modeling, which this study addresses.

\[
Entropy = -\int_{-\pi}^{\pi} \hat{f}(\lambda) \log \hat{f}(\lambda) d\lambda
\]  

Integrating over the spectral density estimate, \(\hat{f}(\lambda)\) and its log value, gives an overall estimate of the importance of each frequency; a high entropy value indicates a highly variable time series, which makes accurate forecasting more difficult.

**Methodology**

This study aims to identify the best forecasting model for hourly building energy consumption, which can vary greatly between hours (see Figure 1). To align with the Infomorphism framework, this study forecasts one time step ahead (i.e. the forecast horizon, \(k\), is equal to one hour ahead), using a sliding window approach to the historical input data to the model. In order to validate the forecasting models proposed, the data is split into 80% utilized for training the models, and 20% utilized for testing the predictive accuracy of the models’ forecasts. This modeling effort utilizes only one input feature: the building’s own historical energy demand data. The advantage of a single input feature is data availability. If other features were included in the forecasting model, historical forecasts of those features would be required, in order to align with the historical training data. Even weather variables (e.g. temperature, wind speed) are difficult to obtain historical forecasts for at the hourly level; data server limitations mean that such information is only stored historically ev-
every 6 hours. An interesting direction for future research would be to study the availability and impact of potentially relevant input features.

Statistical and machine learning methods strive to find an optimal functional form that expresses the relationship between independent input data (X) and a corresponding dependent variable (y). Given the nonlinearity present in the building energy data (see Table 2), a method capable of handling this type of data is required.

Several such machine learning models were trained and evaluated for predictive accuracy. These models include a statistical learning benchmark model - the Persistence method, as well as four state-of-the-art competitive models: Support Vector Regression (SVR), a dense deep learning neural network (DNN), and two tree-based methods - XGBoost and Random Forest.

SVR was selected for evaluation based on its ability to deal with nonlinear partitioning of response hyperplanes and higher dimensions of features (Arrieta-Prieto and Schell (2022)). Previous research on high-frequency, high-volatility data has shown that DNNs have the ability to assign higher weights to temporal dependencies with significantly higher demand-response speeds (Yang and Schell (2022)). Tree-based ensemble methods like XGBoost and Random Forest are well-known for their computational efficiency and ability to handle highly correlated input variables (Friedman (2001)).

**Persistence**

The naïve, or persistence forecast, is calculated in order to assess the forecasting skill of different machine learning models. Given the variability of the building energy data, the median value of the historical data is taken as the forecast value, as in Equation 2 below.

\[
\hat{y}_{i+k|i} = \text{median}(y_i)
\]  

**Support Vector Regression (SVR)**

Support vector machines for regression find the optimal response surface, considering nonlinear transformations, K, of the input independent variables.

\[
Y(X) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i)
\]  

**DNN**

This Dense Neural Network (DNN) refers to a 2 layerfully-connected model.

**XGBoost**

This machine learning method is based on an ensemble of decision trees; including a decision tree in the ensemble model is the result of the gradient boosting algorithm (Friedman (2001)). XGBoost refers to extreme gradient boosting, an implementation which can optimize ensemble construction for any loss function.

\[
\hat{y} = \frac{1}{c} \sum_{i=1}^{c} y_i
\]  

**Random Forest**

The random forest algorithm is based on regression trees, a type of decision tree for continuous dependent variables. Regression trees recursively partition the input data, in order to minimize the predictive error in each of the leaf nodes of the tree. At each leaf node, the dependent variable prediction (\(\hat{y}\)) is given as the average of the observed data (\(y_i\)) within that partition. For a well-built tree, the regression surface at the leaf node is piecewise-constant.

\[
\hat{y} = \frac{1}{c} \sum_{i=1}^{c} y_i
\]

A random forest model grows a specified number of regression trees, B, which are built by random selection of subset \(m < p\) independent variables. A prediction (\(Y_f(X)\)) is made according to Equation 5 by averaging the ensemble of trees \(T(X; \theta_b)\), where \(\theta_b\) characterizes the \(b^{th}\) random forest tree in terms of split variables, cutpoints at each node, and terminal-node values.

\[
Y_f(X) = \frac{1}{B} \sum_{b=1}^{B} T(X; \theta_b)
\]

**Evaluation Metric**

All models developed are evaluated using the same error metric, the mean absolute error (MAE), shown in Equation 6.

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|
\]

Table 3 below shows the MAE results per model and building dataset.

**Results and Discussion**

Table 3 summarizes the results of the model development to forecast hourly building energy consumption. The naïve forecasting benchmark model for time series data, the Persistence model, indeed results in the worst forecasting accuracy, with an error of up to 2.47 kW. The best performing models are the tree-based ensemble models, XGBoost and Random Forest. However, the Random Forest model gives the best predictive accuracy, across all five building data sets.

While a neural network (here, the DNN) is theoretically capable of modeling any functional form, the results show worse performance than the naïve persistence model for building datasets 2 and 5. In the case of building 2, this dataset had the highest variability, so the DNN model overfits to the complex dataset. For both buildings 2 and 5, the hyperparameters of the respective DNN models need to be better optimized, in order for this type of modeling framework to be competitive with the Random Forest. Across all datasets for the DNN model, hyperparameter optimization may help increase the predictive accuracy. However, the need for careful tuning of hyperparameters is a drawback to DNN models compared to Random Forest.
Forests, which achieves better results without any hyper-parameter tuning.

In the context of Infomorphism, these results show that it is possible to accurately predict building energy demand at the hourly level using a Random Forest model. This type of model consistently provides the best predictions across all building datasets in this study. This suggests that only one type of machine learning model needs to be integrated into the Infomorphism framework, which will reduce the overall computational burden.

The forecasting results can be utilized directly in the urban energy network optimization model (Li et al. (2022)), which is embedded within the Infomorphism framework, to predict the demand for Energy Parcels (Li et al. (2022)). The time resolution of the optimization model can be increased to the hourly level without loss of computational tractability, to tackle different planning challenges with increased precision and effectiveness. By integrating the forecasting model results as input to the optimization model, urban planners can optimize renewable integration and develop strategies tailored to specific time scales and study scenarios, leading to a more efficient and resilient urban environment.

### Conclusion

Given the sensitivity of energy demand to real-world events, it is crucial to utilize high-resolution demand prediction models for evaluating and exploring energy-related policies based on different urban scenarios. As the Infomorphism framework allows for the adjustment of supply and demand resolution based on data availability, it enables more precise predictions. This opens up opportunities to expand the framework’s scope beyond long-term planning to include design and planning factors such as scheduling problems in local energy exchange networks and transportation or grid control issues at high-time resolution levels. In addition, moving forward, the integration of the developed forecasting model into the Infomorphism framework can be explored to address issues related to the deployment of advanced building technologies related to renewable energy integration and energy efficiency optimization based on high-resolution settings. As the forecasting model’s resolution increases, the optimization model within Infomorphism can be modified and integrated with new parameters to establish a collaborative workflow for a more collective decision-making process. This comprehensive approach can enable the planning of energy-efficient and renewable energy-integrated cities with an emphasis on equitable energy accessibility in a high-resolution model environment.

### References


### Table 3: Predictive accuracy of hourly building energy consumption, measured by MAE, per dataset and machine learning model. The Random Forest algorithm produced the lowest error, denoted in bold font, across all buildings tested.

<table>
<thead>
<tr>
<th>Dataset No.</th>
<th>Persistence</th>
<th>SVR</th>
<th>DNN</th>
<th>XGBoost</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.477</td>
<td>0.389</td>
<td>0.634</td>
<td>0.344</td>
<td><strong>0.324</strong></td>
</tr>
<tr>
<td>2</td>
<td>1.191</td>
<td>0.843</td>
<td>3.547</td>
<td>0.610</td>
<td><strong>0.584</strong></td>
</tr>
<tr>
<td>3</td>
<td>0.657</td>
<td>0.162</td>
<td>0.504</td>
<td>0.142</td>
<td><strong>0.141</strong></td>
</tr>
<tr>
<td>4</td>
<td>0.261</td>
<td>0.268</td>
<td>0.360</td>
<td>0.238</td>
<td><strong>0.235</strong></td>
</tr>
<tr>
<td>5</td>
<td>1.737</td>
<td>1.161</td>
<td>3.559</td>
<td>0.868</td>
<td><strong>0.787</strong></td>
</tr>
</tbody>
</table>

Mean Absolute Error (MAE) [kW]