OCCUPANCY MODELING USING POPULATION STATISTICS AND MACHINE LEARNING IN THE RESIDENTIAL BUILT ENVIRONMENT

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ORÇUN KORAL İŞERİ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
ARCHITECTURE

SEPTEMBER 2024

Approval of the thesis:

OCCUPANCY MODELING USING POPULATION STATISTICS AND MACHINE LEARNING IN THE RESIDENTIAL BUILT ENVIRONMENT

submitted by **ORÇUN KORAL İŞERİ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Architecture, Middle East Technical University** by,

Prof. Dr. Naci Emre Altun
Dean, **Graduate School of Natural and Applied Sciences** ________________

Assoc. Prof. Dr. Ayşem Berrin Çakmaklı
Head of the Department, **Architecture** ________________

Prof. Dr. İpek Gürsel Dino
Supervisor, **Architecture, METU** ________________

Prof. Dr. Sinan Kalkan
Co-Supervisor, **Computer Engineering, METU** ________________

Examining Committee Members:

Assoc. Prof. Dr. Ayşem Berrin Çakmaklı
Arcihtecture, METU ________________

Prof. Dr. İpek Gürsel Dino
Architecture, METU ________________

Assoc. Prof. Dr. Zeynep Durmuş Arsan
Architecture, IYTE ________________

Prof. Dr. Gülsu Ulukavak Harputlugil
Arcihtecture, Çankaya University ________________

Assoc. Prof. Dr. Bekir Özer Ay
Architecture, METU ________________

Date: 06.09.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name: Orçun Koral İşeri

Signature:

ABSTRACT

OCCUPANCY MODELING USING POPULATION STATISTICS AND MACHINE LEARNING IN THE RESIDENTIAL BUILT ENVIRONMENT

İşeri, Orçun Koral
Doctor of Philosophy, Architecture
Supervisor: Prof. Dr. İpek Gürsel Dino
Co-Supervisor: Prof. Dr. Sinan Kalkan

September 2024, 247 pages

The emphasis of Urban Building Energy Modeling (UBEM) in informing urban design, building assessments, and energy system strategies for understanding the climate change impact on urban environment is challenged by the performance gap which is a disparity between simulated and actual energy demand and can be linked to insufficient modeling of occupant behavior. Thus, this thesis introduces a data-driven occupancy modeling methodology for UBEM. For realistic representation, occupancy modeling necessitates a comprehensive exploration of daily activities and rhythms of occupants. Drawing on data from the population statistics of Italy, this study employs a data-driven approach by incorporating deep learning techniques for holistic occupant modeling with both temporal and non-temporal datasets. By analyzing occupants' profiles, behaviors, and interactions, the study seeks to understand and propose solutions for recognized challenges, leveraging empirical occupancy data that can enhance precision of UBEM.

Keywords: Building Occupancy Modeling, Urban Building Energy Modeling, Residential Energy Demand, Time-series datasets, Data-Driven Models

ÖZ

KONUT YAPILARINDA NÜFUS İSTATİSTİKLERİ VE MAKİNE ÖĞRENİMİ MODELLERİ KULLANILARAK BÖLGE ÖLÇEKLİ KULLANICI DOLULUK MODELLEMESİ

İşeri, Orçun Koral
Doktora, Mimarlık
Tez Yöneticisi: Prof. Dr. İpek Gürsel Dino
Ortak Tez Yöneticisi: Prof. Dr. Sinan Kalkan

Eylül 2024, 247 sayfa

Kentsel Bina Enerji Modellemesinin (KBEM), kentsel çevreye iklim değişikliğinin etkilerini anlamada kentsel tasarım, bina değerlendirmeleri ve enerji sistem stratejilerine katkısının vurgusu, simüle edilen ve gerçek enerji talebi arasındaki performans farkı ile sınanmaktadır. Oluşan bu fark, yetersiz veya yanlış kullanıcı davranışı modellemesine bağlanabilir. Bu tez çalışması, KBEM için veriye dayalı kullanıcı modellemesini geliştirmek için yenilikçi bir metodoloji sunmaktadır. Kullanıcı profillerini, davranışlarını ve etkileşimlerini analiz edip aynı anda modelleyerek KBEM sürecinin performans doğruluğunu artırabilecek deneysel kullanıcı verilerini üretmeyi amaçlamaktadır. İtalya nüfus istatistiklerinden elde edilen verilere dayanarak, bu çalışma, birçok bileşeniyle bütünsel kullanıcı modellemesini modellemek için derin öğrenme tekniklerini dahil eden veriye dayalı bir yaklaşım kullanmaktadır. Çünkü, gerçekçi temsil için, kullanıcı modellemesi, kullanıcıların günlük aktiviteleri ve ritimlerinin kapsamlı bir incelemesini gerektirir.

Anahtar Kelimeler: Kullanıcı Modellemesi, Kentsel Bina Enerji Modellemesi, Konut Enerji İhtiyacı, Zamansal Veriler, Veri-Tabanlı Modelleme

to all my loved ones, may we share many more wonderful days ahead

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Dr. İpek Gürsel Dino, and my co-advisor, Prof. Dr. Sinan Kalkan, for their unwavering support, insightful guidance, and continuous encouragement throughout the entire duration of my research and the writing of this thesis. Their expertise and patience have been invaluable to the completion of this work.

I am profoundly grateful to my family, especially my father and mother, for their unconditional love, support, and understanding. Your faith in me has been a source of strength and motivation.

For their emotional support, I extend my heartfelt thanks to Ataollah Tofigh Kouzehkanani, Buse Ezgi Sökülmez, Gizem Özkan Üstün, Nil Akdede, Sepehr Seyedian Choubi and Şahin Akın. Your encouragement and companionship have been crucial during the ups and downs of this journey.

I also wish to express my appreciation for the guidance and support of my thesis supervisors throughout the research project and writing process. Special thanks go to the committee members and other faculty who provided valuable feedback and advice during the development of my thesis. Your insights have greatly enriched this work. I am grateful for the TUBITAK 2211-A and YÖK 100/2000 scholarships that supported my research. Without this financial support, the completion of this project would not have been possible.

I would also like to acknowledge the contributions of my research collaborators, colleagues, and peers. Your discussions, data sharing, and technical assistance have been instrumental in the progression of my research.

Finally, I would like to acknowledge the institutional resources, facilities, and administrative support comes from METU that have enabled the completion of my thesis research.

TABLE OF CONTENTS

ABSTRACT v
ÖZ vi
ACKNOWLEDGMENTS viii
TABLE OF CONTENTS ix
LIST OF TABLES xiv
LIST OF FIGURES xv
LIST OF ABBREVIATIONS xviii
1 INTRODUCTION 1
1.1 Background of the Study 1
1.1.1 Evaluating Building Energy Models: Towards Sustainability 3
1.1.2 Enhancing UBEM for Effective Residential Energy Management 4
1.1.3 Climate Change and Residential Performance through Occupancy 5
1.1.4 Occupancy: A Critical Factor in Residential Energy Demand 7
1.1.5 Advancing Occupancy Modeling for Accurate UBEM 8
1.1.6 Challenges in Residential Occupancy Modeling 10
1.2 Research Problem 12
1.3 Limitations in Prior Research and Approaches 14
1.4 Contribution of the Thesis 16
1.4.1 Advantages of Occupancy Modeling in Architectural Discipline 19
1.5 Research Questions 20
1.6 Research Aim and Goals 22
1.7 Research Methodology 23

1.7.1 Deficiencies of TUS Datasets of Türkiye ... 24
1.7.2 The Motivation of Selecting the Italian Statistics ... 25
1.7.3 Methodological Steps ... 27
1.8 Dissertation Outline ... 28
2 LITERATURE REVIEW ... 29
2.1 Theoretical Background of Occupancy Modeling ... 30
2.2 UBEM for Residential Building Stock ... 31
2.3 Occupancy Modeling in Urban Residential Building Stock ... 33
2.3.1 Dimensions of Occupancy Modeling ... 33
2.3.2 Components of Occupancy Modeling ... 36
2.4 Studies on the Components of the Occupancy Modeling ... 37
2.5 Modeling Approaches of Occupancy Modeling ... 39
2.5.1 Occupancy Modeling: Basic Approaches ... 41
2.5.2 Occupancy Modeling: Advanced Approaches ... 45
2.5.3 Literature Gaps in the Current Occupancy Modeling Approaches ... 66
2.6 How to Realistically Model the Occupancy ... 67
2.7 Data Collection for the Occupancy Modeling ... 68
2.7.1 Data Quality for the Occupancy Modeling ... 69
2.7.2 Preprocess for Occupancy Modeling ... 69
2.7.3 Data Collection process for Implicit Dataset ... 70
2.7.4 Data Collection Process for Explicit Dataset ... 72
2.8 Occupancy-Related Features ... 75
2.9 Occupancy Integration for UBEM ... 77
3 METHODOLOGY ... 79

3.1 Research Methodology 80
3.1.1 Development of Dataset for Occupancy Modeling 80
3.1.2 Sequential Occupancy Information Modeling 81
3.1.3 Occupancy Integration to Building Energy Modeling...............................81
3.1.4 Analyzing the Results and Outlining Next Steps.......................................82
3.2 Occupancy-Related Datasets 82
3.2.1 The Content of the Dataset, HETUS and Content of TUS 83
3.2.2 The Selected Datasets: TUS and CENSUS 87
3.2.3 Time-Use-Survey Dataset (TUS)...88
3.2.4 The CENSUS Dataset 88
3.2.5 Data Analysis of TUS temporal 90
3.3 Occupancy Dataset for UBEM 92
3.3.1 Occupancy Modeling Related Components..93
3.4 Occupancy Modeling 98
3.4.1 Limitations of Occupancy datasets 100
3.4.2 Augmentation of Sequential Occupancy Information 102
3.4.3 Occupancy Integration: Conversion from TUS Augmented to TUS Energy 123
4 RESULTS 135
4.1 Before Experiments: Pre-processing of Dataset 135
4.1.1 Preprocessing for Census Housing Sub-Dataset, 2011............................136
4.1.2 Preprocessing for TUS Individual Sub-Dataset, 2011 137
4.1.3 Preprocessing for TUS Daily Sub-Dataset, 2011 138
4.1.4 Merging for Temporal Features of TUS Sub-Datasets 140

4.1.5 Preprocessing for TUS Temporal 141
4.2 Experimental Setup: Occupancy Modeling 145
4.3 Experiment 1: Conventional Method (CM): Machine Learning with FE . 147
4.3.1 Hyper-parameter Tuning (HT) of CM 148
4.3.2 Feature Importance for Occupant Demographics 153
4.3.3 Imputation process of missing daily presence schedules with CM 156
4.4 Experiment 2: Stochastic Method (SM): Markov-Chain Sequences 162
4.5 Experiment 3: End-to-End Deep Learning Method (EM): Multi-task 171
4.6 Experiment 3a: EM vs Various Embedding Structures of EM 179
4.6.1 The First Model (ie. Without Embedding Structure) 182
4.6.2 The Second Model (ie. Embedding with Partial Categorical Columns) 183
4.6.3 The Third Model (i.e. Embedding with Full Categorical Columns) 184
4.7 Experiment 3b: EM vs Various RNNs 186
4.7.1 Results of Training Process for the First Model, LSTM 187
4.7.2 Results of Training Process for the Second Model, LSTM 189
5 CONCLUSION 191
5.1 Revisiting the Research Questions 191
5.2 Limitations and Challenges 200
5.2.1 Data Collection 201
5.2.2 Preprocessing 201
5.2.3 Modeling 203
5.3 Main Contributions 204
5.4 Further Works 205
REFERENCES 207

A. Tuning Scores for ML with Feature Engineering 235
B. End-to-End Deep Learning without Feature Engineering, LSTM Model Architectures 237
C. Chi-Square test for Occupant's Presence Schedules: A Statistical Differentiation Test for ML with Feature Engineering Results 238
D. Chi-Square test for Occupant's Presence Schedules: A Statistical Differentiation Test for End to End without Feature Engineering Results 241
E. LSTM Multi-categorical Performance Analysis For Activity Schedule Category Generation In Terms of Precision, Recall And F1-score 242
CURRICULUM VITAE 245

LIST OF TABLES

TABLES

Table 2.1 Deterministic approaches of the occupancy modeling............................43
Table 2.2 Statistical analysis and models of the occupancy modeling in BEM......47
Table 2.3 Probabilistic models of the occupancy modeling....................................49
Table 2.4 Agent-based models of the occupancy modeling....................................55
Table 2.5 Data-driven models of the occupancy modeling.....................................60
Table 2.6 Implicit Datasets for Occupancy Modeling...71
Table 2.7 Explicit Datasets for Occupancy Modeling, others.................................73
Table 2.8 Explicit Datasets for Occupancy Modeling, Questionnaire 74
Table 3.1 A sample from TUS Individual dataset...96
Table 3.2 A statistical description of TUS Daily with only sequential features 96
Table 3.3 A sample from TUS Daily dataset 97
Table 3.4 Set point/back standards for Building Energy Modeling......................127
Table 3.5 Operation Periods of the Conditioning Systems 127
Table 4.1 TUS daily dataset, missingness in terms of Employment Status and Temporal Resolution 139
Table 4.2 TUS temporal, content 139
Table 4.3 Comparison in terms of order of markov-chain with intial state option167
Table 4.4 Comparison in terms of implementing the contextual parameters to the markov-chain model with option of intial state..170

LIST OF FIGURES

FIGURES

Figure 1.1. Six main components of building energy demand 7
Figure 1.2. Occupancy Components for BEM Integration 23
Figure 1.3. Climate Resemblance of Türkiye and Italy 25
Figure 1.4. Flow of the Study 28
Figure 2.1. Dimensions of Occupancy Modeling (Melfi et al., 2011) 35
Figure 2.2. Components of the Occupancy Modeling 36
Figure 2.3. Historical Time Steps of Occupancy Modeling using categories 40
Figure 2.4. Four main modeling techniques (O. Ahmed et al., 2023b) 45
Figure 3.1. Flowchart of the Thesis Study 79
Figure 3.2. Occupancy-related Dataset, Content 82
Figure 3.3. Standardized Question format of HETUS 85
Figure 3.4. Distributions of UBEM-related columns in the CENSUS 89
Figure 3.5. Occupancy Modeling Over Datasets & Steps 98
Figure 3.6. The rationale behind the augmentation of the occupancy dataset 101
Figure 3.7. Methodologies of Occupancy Modeling 102
Figure 3.8. Temporal hierarchy of occupancy dataset 103
Figure 3.9. The main process of ML with Feature Engineering 104
Figure 3.10. The Markov-chain state representations; (left) first order, (middle) second order, (right) higher order (Snodgrass & Ontañón, 2014) 108
Figure 3.11. The transition matrix computation of the activity schedules 111
Figure 3.12. LSTM, model architecture 113
Figure 3.13. Pseudocode of Data Pre-processing of CENSUS, RAW Dataset 115
Figure 3.14. Temporal Encoding for continuous features for LSTM 116
Figure 3.15. Pseudocode of LSTM with Embedding for CENSUS Dataset 118
Figure 3.16. From generated occupancy components to UBEM imputs 123
Figure 3.17. Example of Heating set point E+ structures 126
Figure 3.18. Generation of Ligthing Use schedules using Presence Schedules ... 128

Figure 3.19. Generation of Equipment Use schedules using Presence Schedules 129
Figure 3.20. Metabolic Rate of Human Activities 131
Figure 4.1. Experimental Setup Chart of All Experiments 135
Figure 4.2. The visual proof of the need for imputation, Census Housing 136
Figure 4.3. Missingness ratio of features, after preprocessing, TUS Individual ... 137
Figure 4.4. Merging process without any augmentation for missing schedules ... 140
Figure 4.5. Equalizing the length of all daily schedules...................................... 141
Figure 4.6. Permutation Feature Importance 145
Figure 4.7. Comparisons of the proposed method based on literature review, previous methods 146
Figure 4.8. Application Steps of Conventional Method...................................... 147
Figure 4.9. Tuning progress for the silhouette scores, rest of the plots are in appendix 150
Figure 4.10. Dendrogram, the number of clusters by the levels of linkage 152
Figure 4.11. Feature Importance of the occupant demographic parameters 153
Figure 4.12. Tuning process of t-SNE, best option, right-bottom........................ 155
Figure 4.13. Presence schedule of a household: 4 seasons x 3 weekday category 157
Figure 4.14. Presence schedule of an occupant: 4 seasons x 3 weekday category 158
Figure 4.15. Chi-square test (i.e. p-value) of presence of an occupant 159
Figure 4.16. First-order markov-chain single generated presence schedule example, single state transition 164
Figure 4.17. First-order markov-chain multiple generated presence schedule example, multiple state transition... 165
Figure 4.18. State transition models based on the numebr of the states, transition matrices T (Grewal et al., 2019) 166
Figure 4.19. Activity schedule distribution based on seasonal/weekly variations 172
Figure 4.20. Training curve of LSTM, only periodical features are included in the training.. 173
Figure 4.21. Chi-square results (i.e. p-value) of presence schedule of *occupant x* 175
Figure 4.22. Chi-square (i.e. p-value) of presence schedule of an *occupant y*...... 176

Figure 4.23. Performance metrics for all categories of “Occupant_Activity”...... 178
Figure 4.24. Performance of the first model 183
Figure 4.25. Performance of the second model 184
Figure 4.26. Performance of the third model 185
Figure 4.27. Hyper-Parameter Space visualization with Parallel Coordinate Plot 188
Figure 4.28. Performance of the first model (i.e. embedding with full categorical columns)... 188
Figure 4.29. Hyper-Parameter Space visualization with Parallel Coordinate Plot 190
Figure 4.30. Performance of the second model (i.e. embedding with full categorical columns)... 190

LIST OF ABBREVIATIONS

ABBREVIATIONS

ABM: Agent-based modeling

ACC_ID: Accommodation Identification

ACTs: Activity-related parameters

ACL: Activity Coding List

ANN: Artificial neural network

ARM: Association Rule Mining

ASHRAE: American Society of Heating, Refrigerating and Air-Conditioning Engineers

ATUS: American Time Use Surveys

BEM: Building Energy Modeling

BIM: Building Information Modeling

BN: Batch Normalization

CENSUS: Permanent Census of Population and Housing

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

DHW: Domestic Hot Water

DL: Deep Learning

E+: Energy Plus

EE: Expected Improvement

ENERs: Energy demand related parameters

EPBD: Energy Performance of Buildings Directive

EU: European Union

FE: Feature Engineering

GAN: Generative Adversarial Network

GIS: Geographic Information System

GPS: Global Position System

HABIT: Human And Building Interaction Toolkit

HH_ID: Household Identification number

HHs: Household-related parameters

HETUS: Harmonized European Time Use Surveys

HMM: Hidden Markov Models

HVAC: Heating, Ventilation, and Air Conditioning

ICT: Information and Communication Technology

ID: Identification

IEA: International Energy Agency

IOT: Internet-of-Things

IS: Initial State, for Markov-chain models

ISTAT: Italian National Institute of Statistics

KDE: Kernel Density Estimation

KL: Kullback-Leibler divergence

LSTM: Long short-term memory

MCA: Multiple Correspondence Analysis

MCABM: Markov-Chain Agent-based modeling

ML: Machine Learning

No-MASS: Nottingham Multi-Agent Stochastic Simulation

n-ZEB: Nearly Zero Energy Buildings

OB: Occupant behavior

OCC_ID: Occupant Identification number

OCCs: Occupant-related parameters

OSM: OpenStreetMap

PCA: Principal Component Analysis

PAPI: Pen-and-Paper Personal Interviews

PI: Probability of Improvement

PIR: Passive Infrared Sensor

PFI: Permutation Feature Importance

RESs: Residential-related parameters

RNN(s): Recurrent Neural Networks

SABM: Stochastic agent-based modeling

SCHs: Schedule-related parameters

SHGC: Solar Heat Gain Coefficients

SNN: Spiking Neural Networks

SVR: Support Vector Regression

TPE: Tree-structured Parzen Estimator

TUS: Time-Use Surveys

TUS aug: Time-Use Surveys augmented

TUS clean: Time-Use Surveys cleaned

TUS en: Time-Use Surveys energy

TUS temp: Time-Use Surveys temporal

UCB: Upper Confidence Bound

UBEM: Urban Building Energy Modeling

UK: United Kingdom

t-SNE: t-Distributed Stochastic Neighbor Embedding

CHAPTER 1

INTRODUCTION

Urban Building Energy Models (UBEM) is a tool for assessing energy use in buildings in urban scale. However, these models face significant challenges in accurately estimating energy demand, primarily due to the variability and complexity of occupant behavior. This thesis addresses the critical *performance gap* by introducing a data-driven methodology to occupancy modeling for UBEM. The study aims to model the realistic of occupancy modeling for UBEM performance accuracy. Besides, the proposed model provides a more precise understanding of energy demand patterns in urban residential areas. This introduction outlines the importance of occupancy modeling for UBEM; discusses the challenges of current methodologies; sets the stage for the presentation of a novel approach that improves model accuracy; aids in the pursuit of energy efficiency and sustainability in urban residential buildings.

1.1 Background of the Study

The issue of climate change has been started due to unplanned urbanization, consumption-oriented modern lifestyle, and fossil fuel-based energy generation, leading to significant greenhouse gas emissions and negative environmental impacts worldwide. These changes have resulted in natural disasters, temperature deviations, and disruptions in natural ecosystems (Torriti, 2014). Especially, urban areas demonstrate an observable consequence of rising temperatures and an increased incidence of heat waves (Lomas & Kane, 2013). This challenge presented by urban-specific climate has a direct impact on the energy demand in buildings, emphasizing the requirement for concentrated research in this domain.

Adaptation and mitigation efforts are essential to reverse the negative trends of climate change (Bank, 2010), with the focus on energy efficiency and carbon footprint in urban areas. The Energy Performance of Buildings Directive (EPBD) of the European Union (EU) guides construction and renovation practices for achieving the 2030 and 2050 energy goals, with a focus on eliminating carbon emissions from building stocks by 2050 (EPBD, 2021). The directive highlights the significance of building refurbishment, modernization, and municipal engagement for decarbonization efforts, which depend on the precision and efficiency of energy modeling in urban buildings because urban building stock contribute significant portion of a nation's total energy demand (X. Yang et al., 2022).

To develop effective policies that reduce carbon footprint and improve energy efficiency, it is essential to comprehend the factors that determine energy use in residential buildings, i.e. occupancy behavior, building properties, and daily energy demand patterns. Detailed observations on building performance can enhance the effectiveness, applicability, and efficiency of proposed solutions regarding climate change. For instance, the implementation of retrofit solutions by improving the outer and inner surfaces of the buildings is essential to achieve important reductions in overall energy demand and can prevent to use complex and high-cost technological solutions (Ma et al., 2012).

As we understand the significant impacts of climate change, the role of urban buildings in energy demand becomes increasingly crucial. This highlights the development of Building Energy Modeling (BEM) as a key tool in assessing and improving energy management and efficiency in the urban built environment. Consequently, the progress in the BEM, it is not only framed with technological advancements but also integrating deeper insights into occupancy modeling which has a high impact on the energy use in residential buildings.

1.1.1 Evaluating Building Energy Models: Towards Sustainability

To delve deeper into BEM, its approach from the beginning of the design to retrofit aids stakeholders in analyzing the energy performance of buildings. The evolution of BEM reflects the growing complexity and necessity of accurately modeling energy demand and, it faces challenges, particularly in integrating occupancy modeling.

BEM's approach promotes a holistic view of construction and architectural design (Pan et al., 2023). It helps to stakeholders for analyzing the cumulative impact of diverse systems on the overall building structure, making it especially beneficial in forecasting the energy performance of new constructions. This foresight reduces the likelihood of post-construction design alterations (Pacheco et al., 2012). As the field of building technology evolves, BEM provides architects the capacity to evaluate the effects of emerging systems on energy efficiency, supporting the integration of innovative solutions in building design (Pan et al., 2023).

BEM has undergone significant development over the years. Initially, it was focused on basic thermal calculations, but it has since expanded to encompass a broad range of energy and environmental factors (Ferrando et al., 2020). The BEM framework contains of descriptions for attributes of buildings including geometry, construction materials and systems. The configurations systems basically are HVAC, Domestic heat water (DHW) occupancy schedules, lighting, plug loads and thermostat settings and lastly renewable energy plants. BEM can computer simulations using the local weather data of building. Simulations executes physical equations to calculate thermal loads, system responses and resulting energy demand (Aksoezen et al., 2015; Garg et al., 2017). This includes related metrics such as indoor thermal comfort and total energy demand.

BEM is a fundamental aid in contemporary building design and construction, which provides notable gains in energy efficiency, cost-effectiveness, and environmental impact (T. Hong et al., 2020). Nonetheless, successfully utilizing BEM involves challenges related to software accessibility and technical proficiency, in addition to adapting to the newest technological advancements and sustainability trends in building design. The strategies against climate change and performance evaluation are much more complex than simply assessing individual BEM simulations. They should encompass more by including both static and dynamic neighborhood components, such as citizens.

1.1.2 Enhancing UBEM for Effective Residential Energy Management

The development and enhancement of UBEM is a crucial step forward in addressing the unique energy requirements of urban residential buildings. These models require not only detailed inputs, e.g. GIS data, building characteristics, and occupancy patterns, but also necessitate an understanding of the urban environment's influence on energy demand. The complexity and advanced nature of UBEM highlight the progression from building to urban.

Decarbonization of the buildings starts with a comprehensive analysis on current the conditions of building stock then progressing through for modeling of urban residential building inventory, with a particular emphasis on those with high energy requirements (Ballarini et al., 2011). Precise energy modeling is necessary to implement successful retrofit solutions in terms of climate adaptation and mitigation strategies. Therefore, the creation of more advanced UBEM is crucial for discovering and addressing the unique energy requirements of urban residential buildings.

UBEM requires extensive inputs, including geographic information system (GIS) data, meteorological information, and building characteristics such as materials, year of construction, HVAC system types, occupancy, and equipment usage schedules (Ali et al., 2021). UBEM requires extensive information about individually built structures, but it also requires consideration of the urban environment surrounding each building. Compared to building-scale models, urban-scale models require more complex data sets, and an integration process is needed for synchronization of different data types (C. Wang et al., 2022).

UBEMs are simulation-based models that facilitate multi-building performance analysis by adapting building energy models to neighborhoods, districts, and entire cities (T. Hong et al., 2020; Reinhart & Cerezo Davila, 2016). The complex process of UBEM involves collecting and processing a wide range of datasets from many buildings (Ali et al., 2021; C. Wang et al., 2022). These include energy demand, and environmental impact assessments. These models can be built using a bottom-up, physics-based engineering approach that aggregates and processes individual inputs or using a top-down approach using spreadsheet datasets and applying data driven solutions. Due to this complexity, the computational requirements, and associated costs of UBEM are particularly high. However, advances in technology and computational methods have spurred efforts to automate this process, i.e. development of advanced algorithms and adaptation of machine learning techniques.

1.1.3 Climate Change and Residential Performance through Occupancy

Buildings account are responsible for 30% of the global energy demand, which highlights their critical role in the fight against climate change (Smil, 2017). Especially, residential buildings in urban areas, are the largest energy consumers of all building types because occupant's lifestyles, time spent at home, use of appliances and thermostat settings all play a significant role in energy demand. In addition, the total energy demand increases dramatically during times of extreme temperatures as people prefer to use mechanical cooling or ventilation systems, often leading to an overloaded and insufficient energy infrastructure (Zhang et al., 2022). Thus, this chapter examines how as extreme temperatures exacerbate energy demand and how occupants' lifestyles and behaviors directly influence this demand.

Understanding the nuances of occupancy patterns is important for UBEM. Occupancy in urban areas is often denser, more diverse, and subject to a different pace of life, which has a direct impact on energy demand patterns (T. Hong et al., 2017). The breakdown of energy demand into higher temporal resolution and separate into sub-components is essential, i.e. lighting, appliances, heating, and cooling. The use of each component varies significantly depending on both occupancy and temporal factors, including time of day and season. Recognizing and separating these energy demand categories is critical to accurate modeling.

Conditioning systems, typically the primary energy consumers in residential buildings, have their energy requirements shaped by factors like climate, insulation quality, and HVAC system efficiency (Asim et al., 2022). Additionally, the energy used for lighting depends on the building's design, the availability of natural light, and the efficiency of lighting fixtures. Moreover, household appliances such as refrigerators, washing machines, dryers, and electronic devices contribute significantly to the overall energy demand. All these diverse energy types are intrinsically linked to occupant behaviors, underlining the importance of considering these behaviors in the context of residential energy demand.

Detailed analysis of occupancy patterns in residential settings can lead to significant improvements in energy demand management by altering occupant behaviors. Such insights are crucial not only for enhancing energy efficiency but also for raising public awareness about the urgency of tackling climate change and promoting energy efficiency in urban areas (Tamer et al., 2022). This understanding is vital for advancing decarbonization strategies, laying the foundation for targeted approaches that cater to the unique energy profiles of urban residences. Furthermore, educating occupants on energy-saving practices is key to fostering more conscious and efficient energy use (Kontokosta, 2018).

1.1.4 Occupancy: A Critical Factor in Residential Energy Demand

Occupants key role for residential buildings is most pronounced in residential buildings compared to other building types (Dong et al., 2018). It is one of the six major drivers of building energy demand (O. Ahmed et al., 2023a; Yoshino et al., 2017). Energy use in residential buildings is intrinsically linked to the presence of occupants, as it is primarily for their shelter needs. The significance of occupant demographics, including age, gender, employment status, and education level, profoundly influences energy consumption patterns. These demographic factors contribute to varying lifestyles, routines, and preferences, which in turn affect energy usage. Consequently, the impact of residential occupants on energy demand cannot be overstated.

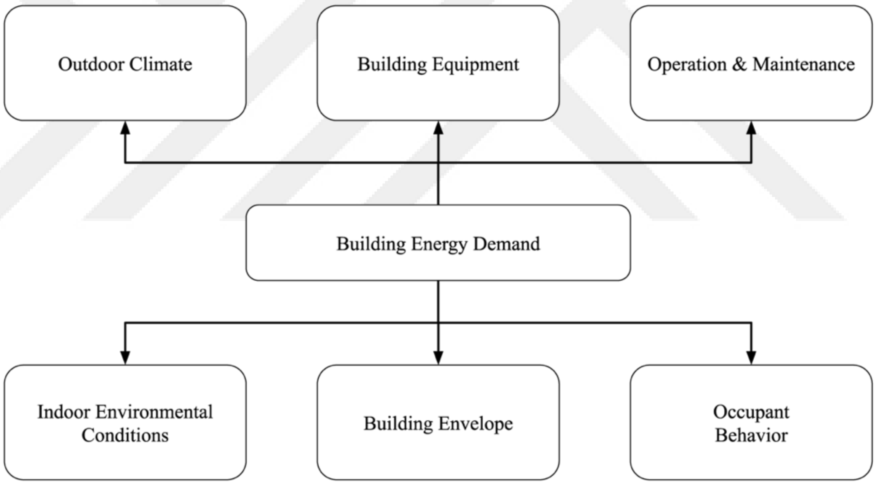

Figure 1.1. Six main components of building energy demand

Occupant profile, presence, habits, and lifestyle choices are the primary drivers of energy use in residential buildings (Santin, 2011). Occupancy density itself shapes overall energy use patterns, with variations in consumption observed based on different occupant behaviors and schedules. The wide range of activities performed by occupants-from cooking and entertainment to working from home-have a direct correlation to specific energy needs. Accurately modeling these activities within

BEM is a significant challenge, as it requires a deep understanding of the complex interplay between different household activities and their respective energy impacts.

Social practice theories examining behavior suggest that activities are often group-oriented and regularly repeated, with interconnected daily routines, e.g. preparation activities before going-out. Accurate modeling of these activities requires identifying these daily, weekly, and seasonal patterns to understand changes in activity and energy use. In other words, people's daily actions, are shaped by fixed schedules for work, meals, and other tasks, lead to predictable energy demand patterns. Therefore, occupancy modeling must account for these consecutive and interconnected daily actions to realistically represent energy demand patterns (Torriti, 2017; Walker, 2014).

To further understand the relationship between occupancy and energy demand in terms of activities, it is important to understand the direct relationship between specific occupant activities and the types of energy demand they generate. Activities as coffee making or using electronics have a pronounced impact on certain categories of energy use. This understanding is critical to developing accurate energy demand forecasts, which require detailed activity data. In addition, distinguishing energy demand at the occupant level is paramount for modeling accuracy. Despite the challenges in data collection and modeling to achieve this granularity, the potential benefits are significant. Occupant-level energy demand analysis can lead to more efficient energy use and personalized energy conservation strategies, ultimately contributing to the broader goals of energy efficiency and sustainability in urban residential contexts.

1.1.5 Advancing Occupancy Modeling for Accurate UBEM

Occupancy modeling is an analytical process that focuses on predicting and analyzing the profile (i.e. demographics), presence, and behavior of occupants in residential buildings (Benezeth et al., 2011; Page et al., 2008). This modeling is critical for estimating energy demand, improving building design and increasing overall energy efficiency. Several key factors are considered when modeling occupancy (Dong et al., 2018; Mosteiro-Romero et al., 2020), e.g. daily, and seasonal variability; diversity of occupant behavior, including different lifestyles, work schedules and personal preferences; household demographics, including household size, household composition, occupant demographics. By anticipating occupancy in the residential buildings, more effective and energy efficient living environments can be created, contributing significantly to the broader goals of sustainable living and climate change mitigation (Kumar, 2007).

Occupant behavior analysis includes how building occupants use energy-consuming equipment and interact with the building environment, such as thermostat settings and window operations. Presence and activity patterns, or schedules, are also critical, outlining when occupants are typically home and active. The various interactions between occupants and their unique preferences for thermal comfort are considered, particularly in residential environments where occupants have more control over the operation of the HVAC system than in commercial buildings (Azar & C. Menassa, 2012; Picco et al., 2014). This detailed analysis of occupant behavior and interactions is essential for developing targeted energy efficiency recommendations in residential buildings.

Occupancy modeling for residential building stock requires the integration of diverse data sources and methodologies to accurately estimate and predict occupancy patterns and their associated energy use (Dabirian et al., 2022). The important step of achieving realistic occupancy modeling lies in the availability of robust and detailed datasets. These datasets are instrumental in capturing the dynamic and complex nature of occupancy patterns, which is a key component in replicating real-world scenarios within building energy models. Furthermore, the application of data-driven methods is essential in translating this data into meaningful insights, thereby ensuring the accuracy and reliability of performance estimations in building energy models. These advancements in occupancy modeling empower urban planners and policymakers with precise tools for managing energy demand, reducing carbon emissions, and guiding sustainable urban development.

This chapter has highlighted the significance of integrating diverse data sources and methodologies for realistic occupancy modeling. Despite these advancements, occupancy modeling for residential buildings is confronted by a range of challenges, particularly in accurately capturing the dynamic and complex nature of occupancy patterns. These challenges are mainly rooted in the stochastic nature of occupant behavior and the limitations of current data collection methods; thus, these are explored in depth in the following chapter.

1.1.6 Challenges in Residential Occupancy Modeling

The occupancy modeling faces several critical challenges and limitations that concludes its effectiveness in building energy simulations (Dong et al., 2018). This chapter delves into these challenges, including the performance gap resulting from oversimplified occupancy assumptions, the complexity of accurately studying stochastic occupant behavior, the need for more granular, occupant-level approaches, data constraints which posing significant challenges, and how data-driven modeling approaches can accurately reflect the intricacies of occupancy patterns.

A significant challenge in occupancy modeling is the performance gap that results from unspecified occupancy assumptions in simulated building energy demand (J. Yang et al., 2016) This gap occurs when the predicted energy use based on simplified occupancy models differs significantly from the actual energy use observed in buildings. This discrepancy is often due to the oversimplified representation of occupant presence and behavior in the modeling process.

Traditionally, occupancy modeling has treated occupant profiles and behavior as independent variables, despite their obvious correlation. Occupant behavior is influenced by a myriad of factors, including age, employment status and lifestyle, which collectively influence energy consumption patterns. Current modeling practices often fail to capture the interrelated nature of these variables, resulting in incomplete or biased representations of occupancy (J. Yang et al., 2016). An integrated modeling approach is needed that includes all aspects of occupant profiles, behaviors, activities, and their interactions to provide a more comprehensive and accurate representation of occupancy patterns.

Occupancy in residential buildings demonstrates a stochastic nature, posing significant challenges to accurate study and modeling (Bosu et al., 2023). This stochasticity is characterized by random and unpredictable patterns of occupant behavior, which significantly complicates the creation of reliable and precise occupancy models. The randomness inherent in how different occupants utilize residential spaces at varying times makes it challenging to predict energy consumption patterns and occupant interactions with building systems. As such, occupancy modeling in residential buildings requires advanced statistical and probabilistic approaches to adequately capture these erratic patterns and translate them into meaningful data for energy modeling.

A major challenge in occupancy modeling is the prevalent use of a household-based approach to determine presence schedules. This approach often fails to capture the true nature of occupancy in a building, as it tends to overlook the individual variability and unique patterns of each occupant within a household (Zhou et al., 2023). To address this, occupancy modeling needs to adopt a more granular, occupant-level approach that considers individual schedules and interactions between occupants. This would allow for a more accurate representation of occupancy patterns and, consequently, a better comprehension of energy use in residential buildings.

Historically, the field of occupancy modeling has been constrained by data limitations, often resorting to assumption-based models rather than relying on actual data. This reliance on assumptions introduces uncertainties about the accuracy of the models, as they may not truly reflect the complex nature of occupancy patterns (Zhou et al., 2023). Modern data-driven approaches, which use actual household responses and behaviors, provide an opportunity to verify and validate the data generated, thereby improving the accuracy and reliability of occupancy models.

Data issues in occupancy modeling present a formidable challenge in the field of building energy simulation (Dabirian et al., 2021; Nejadshamsi et al., 2023). A primary concern is the data availability, or the presence of inconsistent data that is not well-suited for accurate occupancy modeling. Often, multiple sets of data that do not correlate with each other are available, adding complexity to the modeling process.

1.2 Research Problem

The influence of occupancy on the energy demand of residential buildings is profound. A critical factor contributing to the discrepancy between simulated and actual energy demand known as the *performance gap* which is the inadequate integration of occupant behavior in building energy simulations. Unless occupancy presented realistically, this oversight can lead to miscalculations in sizing energy systems, culminating in unnecessary expenditures and diminished efficiency. Consequently, an in-depth and accurate modeling of occupancy patterns is pivotal in bridging this gap.

An accurate assessment of occupancy is essential for a realistic representation of building energy demand. However, the occupancy modeling process can be evaluated as complex problem due to multifaceted nature e.g. occupant profiles, behaviors, presence, daily routines, interactions among occupants, and their engagement with appliances. In many applications of BEM, occupancy is often simplified, generally represented at the household level, or based on assumptions. Thus, more holistic approaches are needed that can include the many components of the occupancy, e.g. schedules as presence, activity and energy usage, demographics as household and occupants.

The precision in occupancy modeling requires diversification to account for individual differences and variations in energy usage patterns, including seasonal, monthly, and weekly fluctuations. Thus, it can be concluded as the temporal resolution increase the accuracy of the modeling improves. This approach to occupancy modeling provides a comprehensive backdrop for understanding and addressing the challenges in residential energy demand within urban environments. Finally, this research aims to enhance the precision of UBEM by leveraging data-driven methodologies and deep learning techniques, analyzing occupant profiles, behaviors, and interactions to yield insights with significant implications for urban planning and policy. The goal is to foster sustainable urban futures through the creation of accurate occupancy models that reflect the true rhythms and activities of urban inhabitants.

Most of the occupancy modeling applications are evaluated based on the assumptions. However, the stochastic nature of the occupant behavior increases the error rate of these assumptions because occupants live with other occupants and interact with the tools, these interactions are hard to model and progress by positioning the assumption-based methodology in the center. This can be more drastic for the residential buildings as the occupant demographics and daily activities vary excessively compared to non-residential buildings. Thus, a realistic information or recordings are needed, and data-driven approaches can provide this due to their top-down nature capturing the overall patterns.

Independent modeling of occupancy often leads to inconsistencies between various occupant components. This arises because these components exert mutual influences because occupant characteristics significantly dictate daily actions or interactions with household appliances. For instance, the tendency to cook varies markedly between adults and children. Moreover, modeling at the household level typically results in a cumulative approach that ignores occupant-level daily activities and interactions. Notably, most studies predominantly focus on generating presence schedules, neglecting the activities that determine occupant presence. For instance, the activity of sleeping, which occurs inside the home, directly implies the presence of the occupant, yet this aspect is often overlooked in simplified models. These oversights highlight the necessity for an integrated approach to occupancy modeling that accurately reflects the complex dynamics of occupant behavior and interaction within residentials.

1.3 Limitations in Prior Research and Approaches

One of the challenges in current methodologies in occupancy modeling is the inherent variability and unpredictability of occupant behavior in the residential unit as the occupant can spend the whole time with various activities and conditions (Yan et al., 2017). Residential occupants do not follow a consistent pattern of presence and activity, making it difficult to generalize or accurately predict their energy use (Baetens & Saelens, 2016; J. Chen et al., 2021; Page et al., 2008). This variability is due to factors as occupant lifestyles, work schedules, and personal preferences, which can vary significantly even within the same building. As a result, models that rely on average or standardized behavior patterns often fail to capture the true diversity of occupant behavior, leading to inaccuracies in energy demand predictions.

Socioeconomic and cultural factors add another layer of complexity to occupancy modeling. These factors influence how different groups of occupant or household use energy in their homes, affecting everything from heating and cooling practices to appliance and lighting use. However, many existing models do not adequately account for these variations, resulting in a one-dimensional view of occupancy (S. Chen et al., 2021). Recognizing and integrating these diverse socioeconomic and cultural influences is essential for a more accurate and comprehensive understanding of residential energy demand (H. B. Gunay et al., 2013).

Integrating occupancy models with other building systems and energy models presents significant challenges (Tuniki et al., 2021). Occupancy is only one aspect of a broader system that includes building materials, weather patterns, and other energy systems. The interplay between these elements is complex, and accurately modeling this interplay is critical for accurate energy demand predictions (Heydarian et al., 2020; Langevin et al., 2015). However, achieving this level of integration is often difficult due to differences in modeling approaches, data formats, and the multidisciplinary nature of the problem.

Occupancy data collection presents its own set of complexities, primarily due to the varied nature of occupant activity and presence. Capturing a comprehensive and accurate data set that reflects the complex aspects of occupant behavior is a daunting task. This task is further complicated by privacy concerns, as detailed data collection can infringe on individual privacy (Padilla-López et al., 2015). Balancing the need for detailed data with ethical considerations of privacy is a critical issue that needs to be addressed in occupancy data collection methodologies.

Another key limitation to accurate occupancy modeling for energy demand forecasting is the availability and quality of data. Often, the data available to researchers and urban planners is either too coarse, incomplete, or outdated. This lack of high-quality, granular data hinders the ability to build detailed models that accurately reflect current occupancy trends. In addition, the data collection methods themselves may be limited due to budget constraints, logistical challenges, or privacy concerns, further limiting the depth and breadth of data that can be collected.

Technological and computational constraints also play a critical role in the limitations of occupancy modeling. Advanced modeling techniques that can accurately represent the complexity of human behavior often require significant computational resources and advanced algorithms. This requirement can be a barrier, especially in large-scale studies or when real-time data processing is required. In addition, integrating such complex models into broader UBEM frameworks can be challenging given the need for compatibility and efficient computation (T. Hong et al., 2017).

Preparing occupancy data for effective integration into the UBEM framework is a challenge that is often underestimated (T. Hong et al., 2020). Occupancy data sets, which are inherently complex and diverse, require careful preprocessing to ensure that they are compatible with and relevant to energy modeling. This process involves cleaning, normalizing, and structuring the data in a way that is consistent with UBEM's modeling objectives. The complexity of this task cannot be understated, as it requires an advanced understanding of both the data and the modeling framework. The challenges of preprocessing are not only technical, but also related to the conceptual alignment of occupancy data with energy modeling parameters.

1.4 Contribution of the Thesis

In this thesis, a novel methodology is developed that combines empirically collected datasets, state-of-the-art data analysis, advanced machine learning models and dynamic simulation techniques. The methodology of this research integrates a blend of empirical data collection, advanced data-driven methods. Deep learning algorithms are applied to time-series data extracted from population statistics to provide a nuanced understanding of how occupant behavior impacts energy demand. This part provides a detailed overview of how this innovative hybrid approach differs from traditional methods of UBEM.

Dependent Occupancy Modeling: This research pioneers a comprehensive approach in occupancy modeling by integrating multiple occupant-related factors which integrates occupant profiles, presence schedules, activity schedules, and occupant-occupant and occupant-environment interactions. While the current literature predominantly focuses on presence models or evaluates occupant parameters separately, this research aims to develop a more holistic approach. By considering the interdependencies between different aspects of occupant behavior, the study targets more accurate representation of occupant dynamics. The resulting model provides a realistic representation of occupant dynamics, advancing our understanding of their impact on energy demand in residential contexts.

Use of Advanced Data-driven Models: This research is the use of advanced data-driven models to classify sequential data, combining with occupant demographics as representing a significant advancement in occupancy modeling. By breaking down energy use into distinct categories such as lighting, appliances, and HVAC systems and examining them on an hourly basis, the study reveals how daily and seasonal variations in occupant activity drive energy demand. This granularity allows for an unprecedented examination of how occupant activities, influenced by variables such as seasonality and weekly changes, drive energy demand patterns. This methodology not only enhances the understanding of the household impact on residential energy demand, but also enables the translation of occupant activities into quantifiable energy demand metrics i.e. binary or percentage levels.

Proposal of Development of an Occupant-Centric Database for Further Steps:

One of the most important contributions of this study is the proposal of a basic framework for the development of a resident-centered database at the national level. While current UBEM datasets focus primarily on energy demand, shifting the focus to occupant behavior can significantly enrich the depth of data and improve our understanding of the interplay between occupant behavior and energy consumption. This framework represents the first step in building such a database. It integrates different datasets through advanced pre-processing techniques to match both temporal and non-temporal data. By using deep learning and population statistics, this approach creates a solid foundation for modeling user behavior and opens up new possibilities for energy demand modeling and other performance-related applications.

Furthermore, the low implementation rates of retrofit scenarios are often attributed to the fact that they do not take human factors into account. An occupant-centric database can bridge this gap by providing important insights into people's preferences and behavior, making retrofit strategies more adaptive and responsive to the challenges of climate change. An occupant-centered approach can not only improve energy efficiency, but also deliver adaptive design and retrofit strategies tailored to the needs of households. It provides insights into space use for more functional and sustainable design, improves indoor environmental quality through better ventilation and lighting, and enables personalized, adaptable solutions, such as smart home systems. By adapting retrofit plans to actual household behavior, this approach could significantly improve the effectiveness of energy efficiency measures. Finally, its scalability and adaptability make it applicable not only to national contexts but also to other countries involved in the Harmonized European Time Use Surveys (HETUS) (eurostat, 2018).

Implications for Occupancy Modeling and Global Application: This research lays the groundwork for automating occupancy modeling, making it easier to replicate in different regional or global settings. The adaptability of the proposed methodology underscores its potential for widespread application, provided the necessary data are available. This aspect of the study not only makes it more applicable to different geographic areas, but also serves as a model for other researchers and professionals to follow and expand upon. The ability to apply this advanced occupancy modeling framework in diverse locations around the world represents a significant contribution to the field and holds the promise of informing and transforming urban energy planning and policy on a global scale.

Unique Data-driven Approach and its Impact: The novel approach presented in this research marks a significant advancement in the field of occupancy modeling. By employing LSTM networks to classify daily activity schedules of occupants, this study introduces a granular level of detail (i.e. occupant level). This multi-categorical classification, incorporating both activity and location, provides a more accurate and dynamic representation of occupant behavior throughout the year. The impact of this refined modeling has three side: firstly, it significantly narrows the "performance gap" in urban building energy models by offering a more precise estimation of energy demand, tailored to daily and seasonal variations in occupant behavior. Secondly, it equips residents and urban planners alike with a deeper understanding of energy demand patterns, fostering more informed decisions in both personal energy usage and urban energy policy development. Lastly, it facilitates a deeper understanding of the connection between energy consumption and occupants daily activities, paving the way for advancements in energy efficiency through targeted interventions.

1.4.1 Advantages of Occupancy Modeling in Architectural Discipline

Occupancy modeling significantly enhances the architecture discipline by enabling the development of more informed and adaptive design and retrofit strategies that cater to the specific needs and behaviors of households. The following key areas demonstrate how occupancy modeling contributes to architectural practice:

- **Architectural Design:** By offering detailed insights into the usage patterns of different building spaces, occupancy modeling allows architects to design environments that maximize functionality and sustainability. Incorporating this data into building designs leads to more adaptable and resilient architectural solutions that better align with occupants' daily routines and interactions within the space.
- **Climate Change and Adaptation Strategies:** Understanding household demographics and lifestyle patterns enables architects to propose retrofit plans that enhance energy efficiency, particularly in response to climate change. For instance, knowledge of how different households use energy for heating, cooling, and appliances allows for the design of interventions that reduce energy demand while maintaining occupant comfort.

- **Indoor Comfort and Mitigation Strategies:** Occupancy modeling also plays a crucial role in improving indoor environmental quality (IEQ). It informs the design of ventilation, lighting, and acoustic systems that promote occupant health and well-being.
- **Near future implementations:** Looking ahead, occupancy modeling will further enable the creation of personalized and adaptive architectural solutions, such as smart home systems that respond to real-time occupancy patterns or flexible spaces that evolve to meet changing household needs.

Finally, the integration of detailed occupancy data into architectural practice not only helps narrow the performance gap in Urban Building Energy Modeling (UBEM) but also leads to the creation of buildings that are more responsive, sustainable, and supportive of their needs of occupants.

1.5 Research Questions

Occupancy and Data-driven Methodology

Main Research Question: How can a data-driven pipeline, integrating both temporal and non-temporal population statistics, be developed to enhance occupancy modeling better representation for urban residential buildings?

Essentials of Occupancy

Sub-Research Question 1: What are the key elements of effective occupancy modeling, and how does incorporating various factors—such as lifestyle, work schedules, personal preferences, and socioeconomic and cultural influences—into a comprehensive model enhance its accuracy?

Occupancy and Dataset

Sub-Research Question 2: What methods and privacy-preserving strategies can be used to balance the need for effective data collection in occupancy modeling while addressing the trade-off between the privacy risks datasets and the generalized representation capacity offered by datasets?

Sub-Research Question 3: What types of information are essential for occupancy datasets to achieve comprehensive occupancy modeling, and what levels of resolution are required regarding data types?

Sub-Research Question 4: How can occupancy modeling be modified for global applicability, and what types of datasets should be chosen to support this objective?

Occupancy Modeling

Sub-Research Question 5: What are the limitations of cluster-based occupancy modeling in accurately reflecting the complex nature of occupant behavior?

Sub-Research Question 6: What are the limitations of assumption-based occupancy modeling in accurately reflecting the stochastic nature of occupant behavior?

Sub-Research Question 7: What are the implications of modeling occupancy at the household level versus the individual occupant level for accuracy of occupancy modeling?

Sub-Research Question 8: What are the key challenges in applying deep learning algorithms to the combination of time-series and time-series data in the context of occupancy modeling, and how can they be addressed?

Sub-Research Question 9: How can deep learning techniques model the complex interactions between occupant profiles and household activities, and how do different modeling approaches compare in capturing this variability in residential UBEM?

Occupancy modeling and UBEM implementation

Sub-Research Question 10: What is the impact of using augmented datasets with seasonal and weekly variations on the precision of machine learning models in predicting occupant behavior.

1.6 Research Aim and Goals

This study presents a novel methodology to improve the accuracy of occupancy modeling in the context of urban residential building stock. The methodology is based on the combination of advanced data preprocessing techniques for population statistics and usage of deep learning models.

The framework consists of a pipeline of transforming population statistics into occupancy modeling information and inputs for the UBEM.

- **Creation of a database occupancy modeling:** Database emphasizes the acquisition, integration, and creation of a database for Occupancy Modeling and the integration for UBEM. This requires a comprehensive database that includes occupancy components as demographics and schedules, building characteristics, meteorological properties, and energy usage patterns.
- **Advanced data-driven occupancy modeling using population statistics:** Proposed approach enhances conventional models by using the combination of data-driven and individual (i.e. occupant) level modeling approach, thereby providing a detailed insight into daily activity and presence routines throughout different seasons and day types. The dataset combines both temporal and non-temporal elements, thereby enabling a comprehensive analysis of occupant behaviors and their association with demographic features.
- **Incorporation of advanced occupancy information for the building energy model:** This study aims to integrate detailed and realistic occupancy components for the building energy modeling for emphasizing the importance of comprehensive occupancy data in enhancing the overall accuracy of urban building energy models.

1.7 Research Methodology

This research takes a hybrid methodological approach, integrating empirical data collection with advanced data-driven and simulation techniques. It is distinguished using deep learning algorithms to analyze time-series data from population statistics, providing a more granular understanding of occupant behavior by applying decomposition-based approach and occupancy impact on energy demand using building energy models.

The decomposition-based method is aiming to break down the aggregate energy usage in residential units into brief time intervals (i.e. 30-minute), i.e. their time spent inside and their daily activities. Also, it explains this total energy demand by linking closely with the demographic details of occupants. Because the characteristics of the occupants shapes the daily actions and the daily actions determine the interaction with for heating, cooling, and lighting, equipment demand and lastly, the internal heat gain by occupants.

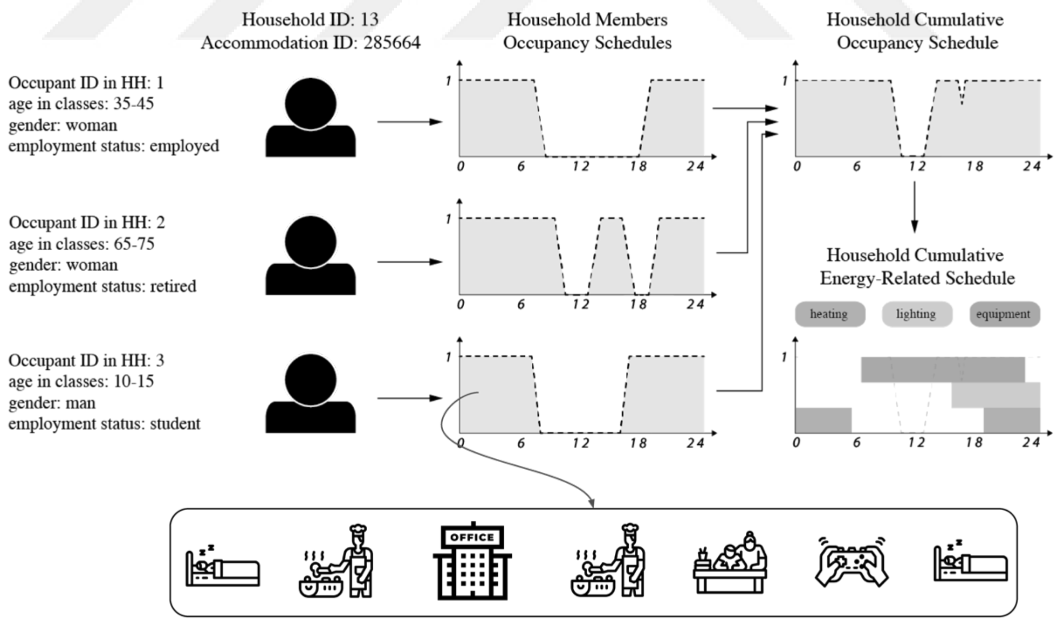

Figure 1.2. Occupancy Components for BEM Integration

Figure 1.2 illustrates an example household, pointing out the sub-components of household modeling. The components of the occupancy model include the schedules that occupants spend at home (SCHs), the daily activities of the occupants (ACTs), household characteristics (HHs), occupant properties (OCCs), finally, the translation of these presence schedules to lighting and equipment usage schedules and equivalents of the activities to metabolic rate-based activity schedules (ENERs).

1.7.1 Deficiencies of TUS Datasets of Türkiye

The TUS data from Türkiye is an extensive survey about residential and household formations in 81 cities of country. However, it provides information publicly up to the city scale. Thus, it is not detailed enough for residential level occupancy modeling without any assumptions (TUIK, 2022). On the other hand, data is lack of detailed about the characteristics of occupants in households, instead focusing more on household features rather than individual occupant profiles. Secondly, TUS data does not align with the HETUS standards (Eurostat, 2019). The discrepancies in data standards and formats pose challenges for applying models to the other regions. However, Türkiye is a candidate country for HETUS, most probably, the subsequent survey can be applicable for occupancy modeling.

Although time use surveys have been conducted in Turkey, the raw data is not publicly available; only statistical analyses are open to everyone. Consequently, TUS dataset from Türkiye was found to be inadequate for detailed occupancy modeling due to its shortcomings. Therefore, this necessitated the exploration of an alternative TUS dataset which can be suited to meet the needs of occupancy modeling for Urban Building Energy Model (UBEM) approach.

1.7.2 The Motivation of Selecting the Italian Statistics

Several key factors are considered in selecting Italian population statistics as the focus of the occupancy modeling and subsequent urban residential building energy modeling. The first one was the climatic resemblance between Turkey and Italy as both countries have coastline for the Mediterranean, especially in the coastline. This similarity offers the potential for wider research applicability and relevance. When Türkiye officially adopted the HETUS standards, the proposed methodology applied to Italy can be transferred to urban areas of Türkiye. The similarity in climate between the two countries could simplify the validation process.

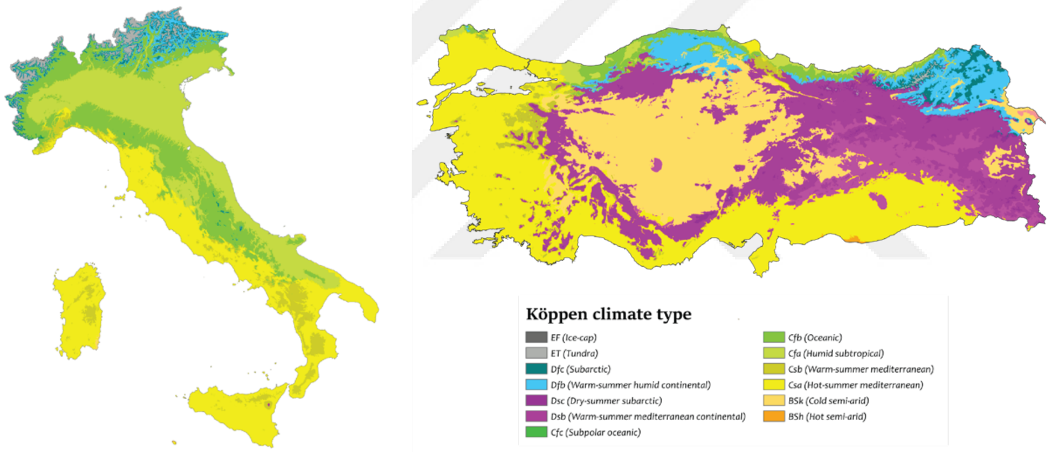

Figure 1.3. Climate Resemblance of Türkiye and Italy

In Figure 1.3, a detailed understanding emerges from the Köppen and ASHRAE climate classifications for exploring the climatic parallels between Italy and Türkiye (Arnfield, 2023; ASHRAE, 2009; Kottek et al., 2006). Both countries are influenced by the Mediterranean climate features. According to the Köppen classification, most of Italy the is categorized as CSA (Hot-summer Mediterranean climate) (Cortellari et al., 2021). These definitions consist of warm, dry summers and mild, wet winters. These features are mirrored along Türkiye' Aegean and Mediterranean coasts (Öztürk et al., 2017). However, Italy's northern regions and elevated areas diverge, exhibiting cooler summers and significant snowfall amount. Türkiye extends beyond these classifications as its broader geographical diversity. The country is encompassing humid subtropical climates along the Black Sea coast and ranging to tundra climates in its eastern extremities. The ASHRAE classification also showcase these similarities and differences. Both Italy and Türkiye's coastal zones under the Zone 2A category which is characterized with hot-humid summers and mild winters. Notably, both countries experience warm, dry summers, with temperatures typically ranging from 20°C to 30°C. Winters are mild, with average temperatures changing between 5°C and 10°C. The total yearly sunshine potential is an average of over 2,500 hours per year. For the geographical features, two countries have significant coastlines along the Mediterranean Sea. This influences their climate in terms of humidity and temperature moderation. Both countries composed of mountainous regions that experience cooler temperatures and higher precipitation.

Second factor is Italy's membership in the Harmonized European Time Use Surveys (HETUS) grants access to standardized, comprehensive time use data. This crucial for robust occupancy modeling. And lastly, Italy's complete and extensive CENSUS and TUS datasets provide a solid data foundation. These datasets cover a broad spectrum of critical data which essential for developing a detailed occupancy model for the urban residential building stock, e.g. occupant and household demographics, physical and thermal properties of the residential building stock, activity and presence schedule information including occupant to occupant interaction.

The comprehensiveness of these datasets not only facilitates precise modeling but also underscores the potential for the models to be applied to similar Mediterranean climates and urban environments.

1.7.3 Methodological Steps

The methodology consists of the steps involved in data collection, processing, and model implementation, highlighting the innovative aspects that distinguish this approach from traditional occupancy modeling approaches. The comprehensive approach employs high-resolution (i.e. 30-min) data-driven occupancy modeling.

The approach simultaneously considers occupant characteristics and household combinations, presence information, activity schedules by considering the occupant-occupant and occupant-appliance interactions.

- **Data Compilation:** Assembling essential parameters for Occupancy Modeling and integration to UBEM simulations. This process incorporates population statistics and Time-Use Surveys (TUS) for occupancy, along with hypothetical building layouts and archetype selection for the determination of thermal properties for BEM simulations.
- **High-Resolution Occupancy Modeling (i.e. Dependent):** Establishing an occupancy model with detailed spatial and temporal resolution, integrating occupancy and household demographics, presence, and activity schedules, and representing seasonal and weekly variations, including household interactions i.e. occupant to occupant and occupant to appliances.
- **Integration into BEM:** Incorporating the classified occupancy models into Building Energy Modeling simulations for residential buildings, enhancing the accuracy of energy demand predictions for simulation models.

Each of these steps represents a critical component in the research methodology, aimed at delivering a more accurate and comprehensive understanding of the relationship between occupancy and energy demand in residential buildings within urban environments.

1.8 Dissertation Outline

This dissertation is divided into five chapters. Each part addresses a different aspect of occupancy modeling in urban residential environments using population statistics and machine learning. The introductory chapter presents the background of the study by discussing the importance of energy demand in urban residential building stock, exploring building energy modeling, and highlighting the role of occupancy in energy demand. The second chapter provides a comprehensive literature review, covering conceptual models related to occupancy modeling for UBEM, and identifying gaps in existing research.

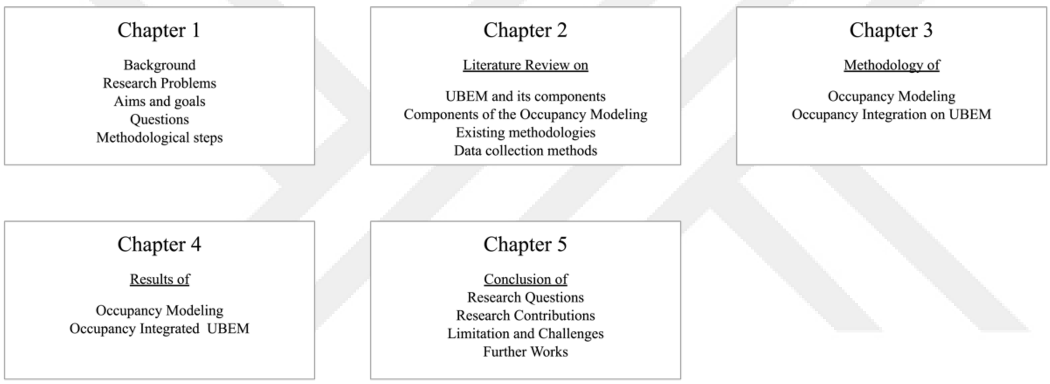

Figure 1.4. Flow of the Study

The third chapter describes the research methodology, detailing the occupancy-related datasets, the rationale for dataset selection, and the process of data pre-processing and integration into building energy models. The final chapter presents the research findings, comparison of performance results of various data-driven methods with proposed occupancy modeling approach. Finally, the dissertation concludes with a summary of key findings and implications for future research at the last conclusion chapter.

CHAPTER 2

LITERATURE REVIEW

The theoretical foundation of occupancy modeling is based on the concept of adaptive ecosystems. This approach evaluates buildings not only as static structures, but as dynamic environments influenced by occupant behavior and characteristics (Balvedi et al., 2018). This becomes more important for the residential buildings as the occupants spend more time and execute variety of activities. Similarly, proposed model of the study aims to capture the complex interplay between occupants' daily rhythms and energy use in terms of seasonal variations, weekday definitions, demographic properties of occupants in residential buildings.

This comprehensive literature review examines the thesis within the existing body of knowledge on occupancy modeling in urban residential building stock. The chapter examines various studies, highlighting the evolution of methodologies and approaches in occupancy modeling, ranging from instance-based, methodological, and resolution-focused strategies to advanced techniques. The chapter not only explains the components and features of occupancy modeling, but also critically evaluates the limitations and gaps identified in previous research. Finally, this chapter underscores the relevance and timeliness of the study and frames the proposed methodology as a response to and advancement of current understanding in the field.

2.1 Theoretical Background of Occupancy Modeling

Occupancy modeling is one of the main contributors that can close the performance gap between simulated and actual energy demand in buildings (Balvedi et al., 2018). Thus, understanding how occupants live and interact with their physical environment is a good starting point (Dong et al., 2022). Behavioral theories present how established habits and daily schedules of occupants affect their presence in particular locations within a building, thereby shaping occupancy patterns (Heydarian et al., 2020). These patterns are crucial for developing models to accurately forecasting or classify the occupancy. While statistical and agent-based models rely on utilizing these existing patterns, data-driven approaches aim to discover them from the data itself. The adoption of diverse theoretical perspectives and a multi-faceted approach is critical to achieving sustainable and efficient occupancy modeling process for built environments. This involves an exhaustive process including resource analysis and in-depth understanding.

The behavior of occupants within urban residential buildings is a key determinant in the overall energy demand for the building stock. Several factors, including the use of appliance, for instance, individual temperature control preferences and daily lighting habits, collectively exert a significant influence on the contours of energy demand within these urban living spaces (Ferrando et al., 2020). Thus, it could be beneficial to delve deep into the daily life of urban dwellers to understand the intricate interplay of these behavioral facets that fundamentally shape the energy profile.

Main Challenges and Solutions

Occupancy modeling for the residential building stock and its integration with energy modeling encounter challenges. One key limitation lies in accurately capturing diverse and dynamic occupant behavior (Schweiker et al., 2017). Existing models frequently rely on standard schedules or statistical averages, overlooking individual variations and contextual factors that significantly affect actual occupancy patterns. Another one is data collection methods often face limitations in terms of scope, accuracy, and privacy. Moreover, the integration of occupancy models with energy modeling tools can be computationally expensive and require specialized expertise.

Thus, occupancy modeling should target to combine several key factors:

- Development of advanced data collection methods to capture high-resolution, context-aware data on occupant behavior.
- Refinement of current theoretical models and exploring new approaches that more accurately represent the complexity of human behavior in various residential settings.
- Incorporating insights from psychology, sociology, and behavioral economics into the modeling framework could enhance its effectiveness. This can be achieved using parameters that are related these subjects.
- Wider adoption and practical implementation can be facilitated by improving the integration of occupancy models with energy modeling tools. This can be achieved by using user-friendly interfaces and standardized data formats and pipelines.

By tackling these challenges, this area could result in significant enhancements in energy efficiency, decreased carbon emissions, and the creation of more sustainable and comfortable urban living environments.

2.2 UBEM for Residential Building Stock

UBEM is an approach to quantifying energy demand in the complex context of urban environments, with a particular focus on residential buildings. This modeling technique is designed to capture the complexities about building properties, thereby providing a holistic understanding of the dynamic energy landscape in urban areas

(Ali et al., 2021). The methodology serves as a critical tool for accurately assessing energy demand and identifying efficiency opportunities, particularly in densely populated urban environments.

UBEM could play a key role in the global effort for the efforts for climate change in terms of modeling and analyzing energy demand patterns of building stock (Ferrando et al., 2020). UBEM provides researchers and policy makers with valuable insights into strategies to reduce the carbon footprint of cities. It highlights the area of energy efficiency in buildings, with a focus on residential structures. The integration of UBEM into climate action plans is increasingly recognized as a cornerstone for the mitigation and adaptation strategies for climate change.

UBEM takes a systems-based approach, considering the built environment as a complicated series of interconnected elements and their consequent impact on each other (Dong et al., 2022). By examining these connections, UBEM can easily analyze energy usage. UBEM integrates concepts of sustainability and resilience, addressing crucial environmental challenges facing urban areas. Through the simulation of various scenarios and interventions, UBEM facilitates the identification of strategies for reducing energy demand, mitigating greenhouse gas emissions, and ensuring a dependable energy supply (Balvedi et al., 2018).

Data-driven methods is a new approach for UBEM, transforming the precision and level of detail in energy demand modeling (Ali et al., 2021). The integration of real-time data from smart meters or wide comprehensive survey studies enables UBEM to create more complex, fine-grained, and up-to-date energy consumption models. Leveraging advanced technologies as machine learning, these approaches enable predictive analysis and the identification of complex demand patterns and shift points (Ferrando et al., 2020). This infusion of data-driven methodologies improves the accuracy and reliability of energy models, overcoming traditional limitations and promoting more informed decision-making.

One of the prevailing trends in sustainable transformation is the gradual integration of renewable energy sources into urban residential energy systems (C. Wang et al., 2022). The adoption of sustainable technologies in these areas is poised to redefine energy consumption paradigms and encourage a transition to greener practices. Expected developments include the establishment of robust smart grids capable of optimizing energy distribution, and the widespread adoption of IoT-enabled homes. This emergence of networked smart homes promises dynamic and efficient energy management, further reducing energy waste and bringing urban energy systems into the realm of sustainability and intelligence.

2.3 Occupancy Modeling in Urban Residential Building Stock

Over the past decade, important steps has been executed in the field of occupancy and occupant behavior modeling, which has significantly improved the analytical capabilities for assessing the performance of buildings, especially energy (Carpino et al., 2017). Initiatives, i.e. Annex 53, have quantitatively presented the impact of occupant activities and behavior on energy demand in both residential and non-residential buildings. Additionally, Annex 66 introduced a comprehensive framework for simulating occupant behavior models and integrating them into building energy performance assessment tools (T. Hong et al., 2017; Zou et al., 2017). Building on the foundation laid by Annex 66, Annex 79 focused on the central role of occupants in informing building design and operational strategies (Yoshino et al., 2017). All these studies are aiming to understand the characteristics and daily life practices of occupants for building energy analysis. Otherwise, erroneous results for building performance analysis may be encountered (Labeodan et al., 2015).

2.3.1 Dimensions of Occupancy Modeling

Occupancy modeling is a multifaceted domain characterized by three critical dimensions (i.e. resolution, granularity): temporal, spatial, and occupancy (Melfi et al., 2011). Temporal resolution ranges from hourly to annual analysis, providing intricate insights into occupant presence and activities over varying timeframes. Spatial resolution delves into the physical context as from broad urban scales down to single thermal zones. Thus, it offers a detailed understanding of where and how occupants interact with their environment. Occupancy resolution encompasses a detailed characterization of occupants, including their number, activities, and movement within the building. This tri-dimensional approach enables a comprehensive analysis of occupant behavior, essential for accurate energy modeling and efficient building design. By integrating these dimensions, occupancy modeling not only captures the complexity of human behavior in residential settings but also enhances the predictive accuracy of energy use and occupant comfort in urban residential buildings as shown in Figure 2.1.

The study of building occupant behavior can be delineated along three critical dimensions (Yan et al., 2015):

- temporal resolution, encompassing various time scales, including annual, weekly, daily, and hourly (Labeodan et al., 2015).
- spatial resolution, encompassing analyses at various scales, from the urban and community levels down to the granularity of individual buildings, zones, and rooms (Yan et al., 2015)
- occupancy resolution, encompassing the characterization of occupants based on presence, number, location, activity, identity and tracking of their movements within the built environment.

Temporal granularity varies and includes annual, seasonal, weekly, daily, and hourly views, all of which offer valuable information on occupant presence and absence. The selection of temporal resolution, spanning from hourly to daily increments, directly impacts the level of detail within occupancy data (Carpino et al., 2017). Hourly data offers a detailed view of occupancy patterns, highlighting fluctuations within broader daily trends. Whereas, daily data displays a general perspective of occupancy peaks, commonly observed during morning and evening hours. Choosing the appropriate temporal resolution requires meticulous consideration of the level of detail desired, weighed against computational complexity. Spatial resolution adds depth to occupancy modeling by revealing the correlation between occupant activity and physical space, particularly as the detail increases to specific residential units (Carpino et al., 2017). Incorporating a longitudinal data collection approach, spanning seasons and years, provides valuable insights into the evolution of occupancy patterns over time (Labeodan et al., 2015). This long-term perspective is crucial in identifying annual trends, seasonal variations, and recurring behaviors within residential buildings, thereby enhancing the understanding of occupancy dynamics and their temporal intricacies.

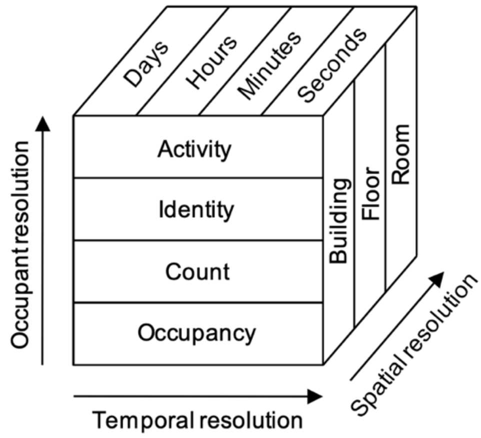

Figure 2.1. Dimensions of Occupancy Modeling (Melfi et al., 2011)

Occupancy modeling in urban building energy models can be executed using two methods. *Space-based* methods examine how the combined behavior of occupants affects specific space categories, depending on the model's spatial granularity (Labeodan et al., 2015). These categories may include archetypal buildings, thermal zones, or functional zones. This approach involves simulating the presence of occupants, controlling space heating, and cooling, and adjusting ventilation rates to meet standard building energy modeling requirements for typical occupancy scenarios. It generates key metrics such as the number of occupants, status of heating systems, and cumulative electrical energy consumption for equipment, all of which are determined by parameters like nominal occupant density, heating setpoint temperatures, and equipment power density.

Person-based methods aim to model the presence, actions of every occupant while considering their unique characteristics and behavioral patterns, including employment status (e.g. full time, part time or unemployed). These models focus on specific details like an individual's control action towards the heating system or appliance usage (Yan et al., 2015). Eventually, they aggregate these actions to achieve outcomes comparable to space-based approaches.

2.3.2 Components of Occupancy Modeling

Occupancy modeling includes information such as the personal characteristics of individuals (occupant profile), the presence of occupants in the home (i.e. occupant presence), the behavior of occupants in the dwelling unit (i.e. occupant behavior), and household characteristics (T. Hong et al., 2017).

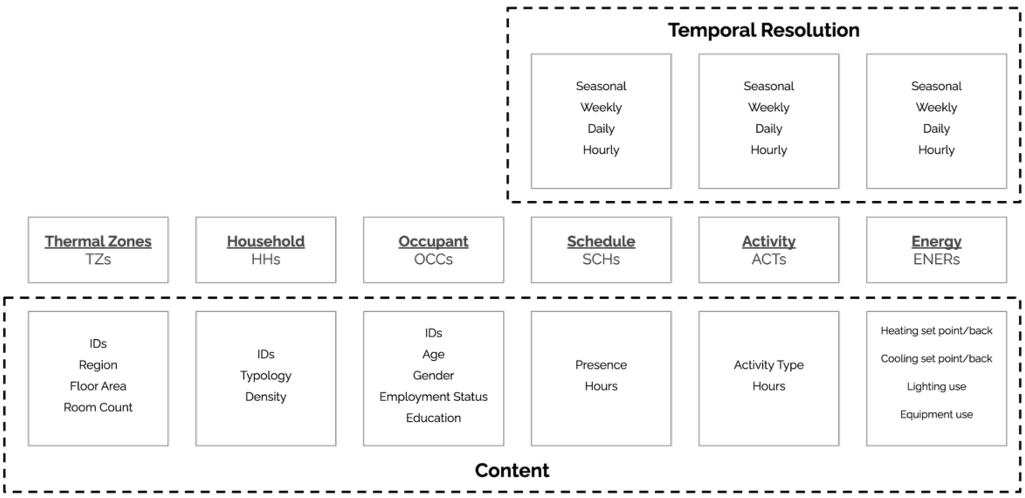

Figure 2.2. Components of the Occupancy Modeling

Here is the general overview:

- *Zone properties*, composed of either physical or thermal parameters, i.e. floor area, total number of rooms in the unit.
- *Occupant demographics*, consisting of variables as age, gender, household composition, employment status, and economic standards.
- The temporal availability of occupants within the residence during a day, captured through *presence schedules*, which are further delineated into hourly, daily, weekly, and monthly intervals.
    - Preferences regarding set points based-one presence of the household members, particularly concerning heating.
    - Encompassing aspects as the intensity of electrical appliance utilization and the specific appliances employed according to presence information, i.e. equipment usage, lighting usage.
- The accumulative behavioral tendencies and preferences of individuals within the household, it is captured with *activity schedules.*
    - Metabolic rate-based activity schedules while considering the daily activities of the occupants, e.g. sleeping, cooking etc.
- The complex interactions among household members, i.e. occupant, offering insights into how occupants engage with one another within the residential milieu i.e. *lighting use schedule*, *equipment use schedule*.

2.4 Studies on the Components of the Occupancy Modeling

Occupancy modeling encompasses several key components that together shape the understanding of how occupants interact with their built environment, both for indoor and outdoor interactions. These components provide a comprehensive framework for characterizing occupant behavior in urban residential buildings (Gaetani et al., 2016) i.e. occupant profile, presence schedule for indoor environment, daily activity schedules, household interactions and energy conversion of the daily activities.

Occupant profiles are the fundamental element of occupancy modeling which include a set of individual attributes e.g. age, gender, , employment status, educational background, household composition and socio-economic factors (Yan et al., 2017). By analyzing these profiles, occupancy models can be tailored to specific occupant groups (i.e. archetype characterization for occupants), enabling more accurate predictions of behavior, energy demand and thermal comfort preferences. In addition, understanding occupant profiles helps to design targeted interventions and policies for energy efficiency and sustainability initiatives that address the unique needs and preferences of different demographic segments (H. B. Gunay et al., 2013).

Occupancy presence information provides a temporal perspective through which to examine the flow of occupants within a building or specific spaces (Tuniki et al., 2021). These patterns cover a range of temporal resolutions, from annual and seasonal variations to weekly, daily, and hourly rhythms, revealing when occupants are typically present or absent. These insights are invaluable for optimizing building operations, energy management and comfort control. Accurate modeling of occupancy patterns enables building systems to anticipate occupancy and adjust lighting, heating, cooling, and ventilation accordingly. Various research are conducted for the prediction of the occupancy presence in the building environment (S. Chen et al., 2021).

Occupancy activity patterns are concerned with the actions and behaviors of occupants while they are in the built environment. These include a wide range of activities, from using electrical appliances, adjusting thermostats, and operating lighting systems to other energy-related actions (S. Chen et al., 2021. Activity patterns influence the load profiles of buildings, affecting peak energy demand and overall demand. By aligning occupant behavior with energy management, residential buildings can achieve greater sustainability and resource conservation.

Occupant interactions within a residential environment are complex and multifaceted (Yan et al., 2017). These interactions involve social dynamics, joint decision making and the influence of one occupant on another. Understanding how individuals interact within a household or dwelling unit is crucial for modeling collective actions, especially in shared spaces. Interactions can influence energy-related decisions, e.g. setting thermostat preferences, operating appliances, and managing lighting. These social dynamics underscore the importance of considering the interconnectedness of occupants when building occupancy models. Capturing these interactions improves the accuracy of predictions and contributes to more effective energy management strategies in urban residential buildings. There is much research that exist for classifying the interactions for urban residential building stock (Fabi et al., 2015; Langevin et al., 2015).

Energy-related activities cover a wide range of behaviors that have a direct impact on energy demand (H. B. Gunay et al., 2013). These activities include using electrical appliances, adjusting heating, and cooling systems, controlling lighting, and other actions that have a direct impact on energy demand. The intensity and frequency of energy-related activities can vary widely between occupants and over time. By incorporating these behaviors into occupancy models, urban residential building stakeholders can develop targeted energy conservation strategies, implement load-shifting practices, and promote energy-efficient habits among occupants.

2.5 Modeling Approaches of Occupancy Modeling

The concept of occupancy encompasses the scheduling of occupant profile, presence, activity within a defined space, while occupant behavior refers to the interactions between occupants and various devices, e.g. appliances, lighting, and thermostats (Labeodan et al., 2015). Modeling process compose of spectrum of methodologies designed to capture and represent the demographics, presence, and behavior of occupants within built environments, as shown in Figure 2.3. It can be ignored with adoption of fixed schedules for presence and activity information or preferring to implement assumption-based strategies with deterministic characters or choosing to increase the detail by advanced modeling approaches.

Realistic occupancy is essential to mitigate the "performance gap”. Initially, occupancy in BEM was approached as a static component and is implement either with default schedules or constant people density and researchers were given most of their focus on building physical properties (Yan et al., 2017). Over time, researchers started to view occupancy as dynamic problem and they introduced various modeling techniques from agent-based to Markov chain models (Dong et al., 2021). As the field evolved, researchers began considering occupant demographics using clustering machine learning models and mostly focusing on household numbers, age, and employment status (El Kontar & Rakha, 2018). Subsequently, agent-based models with assumption-based approaches are introduced to simulate occupant interactions, particularly in residential settings (Y. Chen et al., 2018).

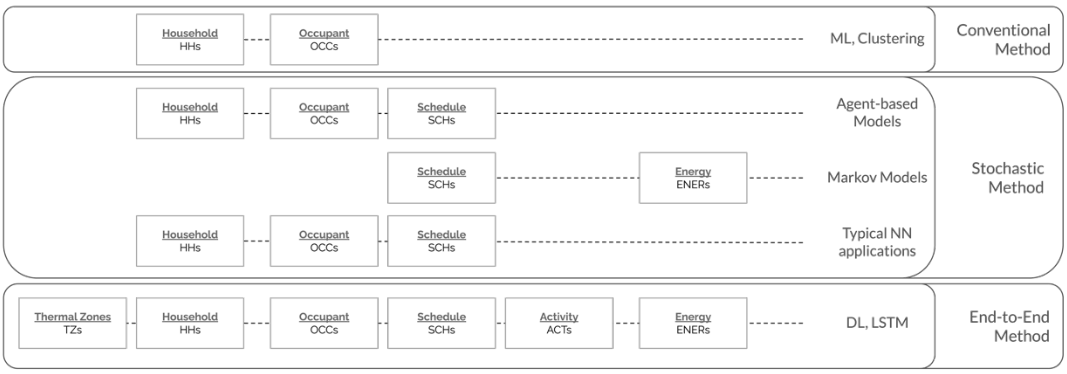

Figure 2.3. Historical Time Steps of Occupancy Modeling using categories

Agent-based models have limitations in simulating multi-occupant activities, focusing mainly on single-occupant interactions with energy-demanding equipment. Non-sequential demographic data was the only input data. Then with the Markov models, researchers start to predict presence schedules based on probabilistic principles (Page et al., 2008). These models aimed to understand how occupancy impacts energy demand. Unfortunately, Markov chain models (MC), based on the premise that activities evolve over time and depend solely on previous states, failed to capture the full spectrum of occupancy variance, especially from an urban perspective. Few researchers tried to model occupant activities by decreasing in the total number of activities modeled which was a resolution problem in the beginning (Wilke et al., 2013).

As a result, a data-driven approach using TUS emerged as a potential solution for detailed occupancy modeling based on real data collection instead assumptions. Nonetheless, many studies preferred to use MC models with TUS datasets and their results have shown inefficiencies, especially in capturing the long-term dependencies of residential occupancy (Richardson et al., 2008).

Historically, before the advent of deep learning, models predominantly used non-sequential data due to computational limitations. The combination of sequential and non-sequential data was not explored extensively due to the significant increase in computational demands. Then deep learning models were introduced for sequential information modeling for occupancy however, still the complete occupancy modeling was not available in the field. Therefore, this thesis proposes the use of Long Short-Term Memory (LSTM) models, which are adept at capturing long-term dependencies in mobility behavior, combined with occupancy demographics. This approach utilizes both TUS and CENSUS datasets, aiming to simultaneously generate activity and presence schedules with enhanced accuracy and detail.

2.5.1 Occupancy Modeling: Basic Approaches

Fixed schedule models are based on pre-established occupancy patterns and schedules by standards (Yan et al., 2017). These models typically adhere to standard assumptions about when and how rooms are traditionally occupied. While they offer simplicity, these models may lack the adaptability needed to account for real-world variations (H. B. Gunay et al., 2013). For example, the ASHRAE standard 90.1 specifies uniform hourly schedules for weekdays and weekends, covering occupancy, plug loads, lighting, infiltration, and thermostat settings for different types of buildings. These are typically used when precise information about actual occupancy and operational schedules is lacking. The schedules are consistent throughout a standard yearly simulation, with variations largely confined to different times of the day. Moreover, they are applied uniformly across all residential models, disregarding any unique characteristics or traits of the occupants (ASHRAE, 2013).

Prior to the emergence of advanced modeling techniques, much research focused on deterministic approaches for occupancy modeling in the context of improving urban building energy efficiency. These studies often utilized pre-defined schedules established by various institutions as a foundational element for their models (ASHRAE, 2013). The recent literature, however, explores a diverse range of methodologies beyond these traditional deterministic approaches. Besides, it's important to note that before 2010, most research in this field primarily relied on deterministic approaches for occupancy modeling in urban building energy efficiency studies. The emergence of advanced techniques has since broadened the scope of methodologies employed.

Deterministic models are based on explicit rules and algorithms to simulate occupant behavior (Yan et al., 2017). These models follow predefined rules and assume that occupant actions will follow predictable patterns. They are useful in scenarios where occupant actions can be predicted with a high degree of certainty. These models are conceptually simple as they operate with clear rules and assumptions. However, they may have limitations in capturing the full range of real-world variations in occupant behavior (Yan et al., 2017).

Table 2.1 Deterministic approaches of the occupancy modeling

Model	*Approach*	*Building Type*	*Data type*	*Output*
Review (Happle et al., 2018)	Categorized occupant behavior modeling methods into deterministic or stochastic and person-based or space-based types.	Various	None	Highlighted mixed-use models' limitations, suggesting improvements.
Review (Mosteiro-Romero et al., 2020)	Assessed occupancy modeling approaches as either deterministic or stochastic in nature.	Institutional	Measurement data	Revealed occupancy, energy demand discrepancies.
Review (Mitra et al., 2020)	Employed ATUS data to establish typical occupancy patterns in U.S. homes, considering household composition and age of occupants.	residential	Time-Use-Survey	Developed nuanced household occupancy profiles.
Review (Happle et al., 2017)	Implemented deterministic occupancy in conjunction with detailed ventilation analysis.	Various	Measurement data	Analyzed district-wide energy demand variations.
Review (Duarte et al., 2015)	Analyzed the effects of ASHRAE"s recommended occupancy diversity factors versus actual data in comprehensive building energy simulations.	office	ASHRAE references and measured data	Showed significant ASHRAE occupancy energy differences.
Deterministic (Yan et al., 2017)	IEA EBC Annex 66 devised a framework for OB research, merging various disciplines and creating quantitative models and co-simulation tools for BPS.	Various	collected data	Formulated occupant behavior modeling approaches.
Deterministic (Lu et al., 2017)	Suggested a robust method to size renewable energy systems in nZEBs, integrating uncertainties in resources and demand through Monte Carlo simulation.	mix	Measurement data	Explored energy balance, mismatch ratio link.
Deterministic (ASHRAE, 2013)	Set a standard for commercial building energy codes in the U.S. and globally for over 35 years.	non-residential	collected data	Estimated occupancy duration in buildings.
Deterministic (SIA, 2006)	Outlined three key roles: formulating calculation assumptions for premises, setting conditions for comfort and lighting, and defining power and energy demand values for different installations.	Various	collected data	Defined building occupancy duration estimates.
Deterministic (Armstrong et al., 2009)	Created in-depth electrical demand profiles for Canadian households using a comprehensive approach, including appliance usage and occupancy trends.	residential	Measurement data	Created varied household energy profiles.
Deterministic (Buso et al., 2015)	Explored the influence of different occupant behavior patterns on building envelope design robustness.	office	collected data	-
Deterministic (Mahdavi & Tahmasebi, 2015)	Compared two probabilistic and one original non-probabilistic model in predicting short-term occupancy trends based on monitored data.	office	None	Predicted workplace occupancy against actuals.
Deterministic (Davis & Nutter, 2010)	Formulated distinct occupancy schedules for various university building types through direct occupancy measurements.	Institutional	Measurement data	Established building occupancy duration schedules.

Happle et al. (2018) reviewed approaches of occupant behavior modeling. They classified the models as deterministic or stochastic revealing limitations in current models, especially in mixed-use districts, and the potential of more accurate predictions with advanced stochastic person-based models. Mosteiro-Romero et al. (2020) evaluated occupancy modeling approaches in institutional buildings and introduced a novel population-based approach (PopAp), revealing significant deviations in predicted occupancies and energy demands compared to standard methods. Mitra et al. (2020) preferred to use American Time Use Survey (ATUS) to generate occupancy schedules for U.S. residential buildings, uncovering significant differences in occupancy schedules compared to those used in current energy modeling methods. Happle et al. (2018) used deterministic-based occupancy schedules and different ventilation calculation techniques on heating and cooling demand forecasts. Lastly, Duarte et al. (2015) compared the impact of using ASHRAE recommended occupancy diversity factors versus collected occupancy data in office buildings. They demonstrated significant variations in total energy demand.

In the specific area of deterministic approaches to occupancy modeling, several studies have made impactful contributions. Kotireddy et al. (2017) examined the impact of occupant behavior and climate change uncertainties on low energy building performance, comparing various robustness assessment methods and highlighting their implications for decision-making. Yan et al. (2015) under IEA EBC Annex 66, developed a comprehensive framework for occupant behavior (OB) research, overcoming challenges in collecting accurate OB data and integrating multidisciplinary approaches. Lu et al. (2017) proposed a design approach for sizing renewable energy systems in nearly zero energy buildings (nZEB), demonstrating its effectiveness in the Hong Kong Zero Carbon Building case study. ASHRAE (2013) was important to set energy standards for non-residential buildings, providing occupancy schedules and duration estimates in the USA. SIA (2006) defined calculation hypotheses and conditions for comfort and energy demand in various building installations in Europe. Armstrong et al. (2009) developed detailed electrical demand profiles of Canadian residentials using a bottom-up approach, addressing limitations in understanding baseloads and small appliance/lighting loads. Buso et al. (2015) investigated how occupant behavior patterns affect the robustness of different building envelope designs. Mahdavi & Tahmasebi (2015) evaluated existing probabilistic models against a non-probabilistic model for predicting short-term occupancy patterns. They have showcased that the probabilistic models are more effective to capture the occupant behavior. Lastly, Davis & Nutter (2010) developed occupancy schedules for institutional buildings. They characterized distinct occupancy profiles for various building types.

2.5.2 Occupancy Modeling: Advanced Approaches

Several advanced methods for occupancy modeling have been formulated by researchers employing diverse techniques and have been grouped into distinct categories.

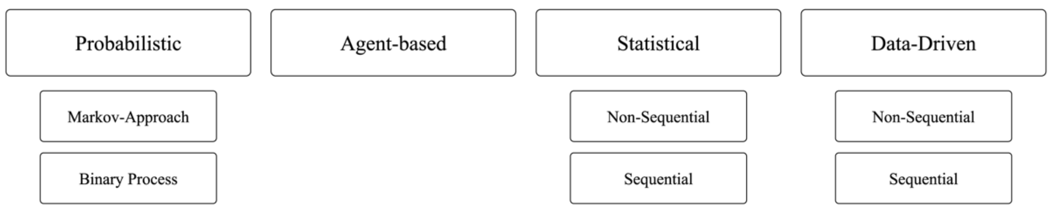

Figure 2.4. Four main modeling techniques (O. Ahmed et al., 2023b)

Agent-based models explore the presence and actions of individual occupant behavior by treating each occupant as an autonomous agent (Gaetani et al., 2016). These agents have different characteristics, preferences, and behaviors. This modeling approach reflects the heterogeneity observed among occupants and their interactions within complex scenarios. Agent-based models excel at capturing the nuanced dynamics of occupant behavior and offer a high degree of realism. They are particularly well suited to scenarios where diverse occupant behaviors and interactions play a key role.

It can be challenging to compose accurate prediction models due to their complicated characteristics and stochastic characteristics nature. *Probabilistic models* introduce an element of randomness and probability into occupancy modeling (J. Chen et al., 2021). They recognize the inherent uncertainties associated with occupant behavior. Stochastic models excel in scenarios where occupant actions are less predictable or exhibit variability over time. By incorporating probabilistic elements, these models provide a means of capturing the dynamic and uncertain nature of occupant behavior, thereby enhancing their applicability in real-world contexts (H. B. Gunay et al., 2013).

Data-driven models use real-time information from multiple channels, such as sensors, smart meters, or surveys, to infer occupancy patterns. Using machine learning and statistical techniques, these models analyze empirical data to accurately predict and understand occupant behavior (Buttitta et al., 2019; Panchabikesan et al., 2021a). They provide a data-centric perspective, enabling the construction of accurate representations of actual occupant actions. Their robustness is evident in their adaptability and ability to learn from data, making them highly effective in scenarios where behavior patterns may evolve or manifest complexity. These models are particularly useful in the construction of high temporal resolution residential occupancy models, facilitating the generation of detailed heating load profiles for occupancy-integrated archetypes (Yan et al., 2017).

2.5.2.1 Statistical Modeling

Statistical modeling plays a crucial role in understanding how occupants' behavior affects building energy use. It does this by establishing mathematical relationships between occupant actions, energy demand of buildings data and various environmental factors e.g. temperature and CO_2 levels. These models help us see how occupant behavior interacts with changing conditions inside and outside the building. They are especially useful for identifying the key factors that influence occupancy patterns. A major advantage of statistical modeling is its ability to analyze how different factors impact occupancy, providing valuable insights for improving building performance. However, a key limitation is that it cannot fully capture the inherent randomness of individual behavior. This highlights the difficulty of predicting occupancy with perfect accuracy and emphasizes the need for more advanced methods that can account for the unpredictable nature of human activities buildings.

Table 2.2 Statistical analysis and models of the occupancy modeling in BEM

Model	*Approach*	*Building Type*	*Data type*	*Output*
Statistics (Santin, 2011)	Analyzed Dutch dwelling occupant behavior to identify energy-related user profiles	residential	Measurement data	Identified five distinct behavioral patterns.
Statistics (D"Oca et al., 2014)	Methodology combines probabilistic models for energy simulations in Dutch dwellings.	residential	Measurement data	Probabilistic heating energy consumption distributions.
Statistics (Fabi et al., 2013)	Probabilistic approach models occupant behaviors impact on residential energy use.	residential	Measurement data	Energy, environment quality probability distributions.
Statistics (Xu et al., 2009)	Studied consumer behaviors effect on energy in Tianjin"s district heating.	residential	Measurement data	Consumer behavior influences water, energy.
machine learning, logistic regression (McLoughlin et al., 2012)	Irish dwellings electricity use analyzed using smart metering and regression models.	residential	Measurement data	Estimates occupancy duration, presence schedule.
machine learning, logistic regression (Yao, 2018)	Methodology models stochastic behavior for air conditioning usage predictions in buildings.	residential	Measurement data	Models stochastic air conditioning behavior.
machine learning, logistic regression (Fabi et al., 2015)	Stochastic behavior for complete HVAC usage	residential	Measurement data	Models predict window states accurately.
machine learning, neural network (Anand et al., 2019)	Developed "Energy-use per person" model using regression and neural networks.	Institutional	Measurement data	Estimates occupancy duration, presence schedule.

Statistical Occupancy Modeling with Statistical Analysis

Recent research in occupancy modeling for building energy demand modeling has yielded significant insights, as shown in Table 2.2. Santin (2011) focused on Dutch dwellings, identifying five Behavioral Patterns (i.e., "spenders", "affluent-cool", “conscious–warm”, “comfort”, and “convenience”) through statistical analysis, but faced limitations due to small sample sizes and the focusing only the new constructions. D’Oca et al. (2014) introduced a probabilistic model combining window opening and thermostat adjustments, highlighting occupant behavior’s considerable impact on energy use. However, the findings require additional confirmation in various climate conditions. Fabi et al. (2015) proposed a probabilistic approach specifically for window opening behavior in residential buildings, creating four models that revealed significant variations in air change rates and space heating demand. Finally, Xu et al. (2009) explored occupant behavior in district heating systems in Tianjin, China, uncovering the influence of behavior on hydraulic performance and energy demand, and identifying a potential 10% energy savings with heat metering systems. However, current models struggle to fully capture the complexity of occupancy dynamics, limiting their ability to generate realistic representations and hindering the development of advanced occupancy-related modeling approaches.

Statistical Occupancy Modeling with Machine Learning (ML)

Machine learning techniques have been increasingly employed to enhance the accuracy of statistical occupancy modeling in building energy demand studies. (McLoughlin et al., 2012) analyzed electricity demand in Irish dwellings using linear regression models, focusing on socio-economic variables and electricity parameters to estimate occupancy duration. However, their approach has limitations due to the exclusion of specific appliances and a low explanatory power in some models. Yao (2018) proposed a stochastic model for air conditioning usage using logistic regression, which was integrated into the EnergyPlus engine to predict usage probabilities and set-points. This model, while improving simulation accuracy, did not account for behavior variability across different locations or occasional events. Similarly, Fabi et al. (2015) developed logistic regression models to predict window opening and closing behavior, revealing correlations with environmental factors like CO_2 concentration and illumination. Their model, however, showed significant errors in predicting opening duration. Anand et al. (2019) took a different approach by developing a "Energy-use per person (K)" parameter using Deep Neural Network algorithms, demonstrating that increased occupancy leads to nonlinear decreases in energy use per person.

2.5.2.2 Probabilistic Modeling

In the realm of occupancy modeling, Markov models are utilized to predict occupancy behavior (OB) by leveraging the system's current state to forecast its subsequent state(s) (Yan et al., 2015). The number of states can vary depending on the model's order, with first order and higher-order models being the most common examples. This approach is advantageous when the transition probabilities between states are known or can be estimated.

Table 2.3 Probabilistic models of the occupancy modeling

Model	*Approach*	*Building Type*	*Data type*	*Output*
markov chain with bernoulli process (Haldi & Robinson, 2011)	CitySim Tool Integration: Simulates occupant behavior and window control effects using advanced models within the CitySim tool.	office	collected data in EPFL	Presence Schedule
markov chain with bernoulli (i.e. discrete) process (B. Gunay et al., 2015)	Modeling Schemes for Occupancy and Lighting: Presents various models from literature for occupant presence and lighting consumption in building performance simulations.	office	collected data	Presence Schedule
markov chain with bernoulli (i.e. discrete) process (Haldi & Robinson, 2009)	Hybrid Modeling Approaches: Develops and tests models, including logistic distributions and Markov chains, for comprehensive stochastic behavior analysis.	office	collected data in EPFL	Presence Schedule
Markov-Chain, first-order (Erickson et al., 2011)	New Occupancy Estimation System: Integrates a novel occupancy estimation method into HVAC systems for refined demand control.	office	derived from sensor data	Presence Schedule
Markov-Chain, first-order (Page et al., 2008)	Markov Chain for Time Series Data: Utilizes a Markov chain model to create time series data representing occupant presence states.	office	collected data in the offices of the LESO	Presence Schedule
Markov-Chain, first-order (Andersen et al., 2014)	Inhomogeneous Markov Chains with Enhancements: Employs inhomogeneous Markov chains with time-of-day adjustments and observation filters.	office	collected data in the offices in San Francisco	Presence Schedule
Markov-Chain, first-order (Wilke et al., 2013)	Ten-Minute Interval Occupancy Data Model: Generates detailed occupancy data at ten-minute intervals, differentiating weekdays from weekends.	residential	Time-Use-Survey	Presence Schedule

Table 2.3 Probabilistic models of the occupancy modeling (continued)

Markov-Chain, first-order + statistical measurements (Wilke et al., 2013)	Time-Dependent Activity Probabilities: Models probabilities of home presence, activity initiation, and activity duration.	residential	Time-Use-Survey	Presence Schedule
Markov-Chain, semi-order + Monte Carlo Simulations (Bottaccioli et al., 2019)	Residential Load Profile Development: Focuses on creating detailed profiles for residential energy load.	residential	Time-Use-Survey	Presence Schedule
Markov-Chain, higher-order (Flett & Kelly, 2016)	Occupancy Model by Type: Develops an occupancy model that differentiates based on occupant and household types.	residential	Time-Use-Survey	Presence Schedule
Markov-Chain, higher-order (Wolf et al., 2019)	Inhomogeneous Hidden Markov Model: Applies an inhomogeneous hidden Markov model for nuanced occupancy simulation.	residential	Time-Use-Survey	Presence Schedule
Markov-Chain, higher-order (Widén et al., 2009)	Non-Homogeneous Markov Chain Application: Utilizes a non-homogeneous Markov chain for residential energy modeling.	residential	Time-Use-Survey	Presence Schedule
Hidden Markov Model (Liisberg et al., 2016)	HMM for Classifying Occupant Behavior: Uses Hidden Markov Models to analyze residential electricity consumption for occupant behavior classification.	residential	an apartment building in Catalonia	Presence Schedule
Hidden Markov Model (Alfalah et al., 2023)	Training HMM with Diverse Occupancy Attributes: Trains HMM under various scenarios using attributes like day, week, and month for optimal occupancy data modeling.	education	collected data	Presence Schedule
Hidden Markov Model (Virote & Neves-silva, 2012)	Rule-Based Occupancy Modeling: Implements a rule-based approach for predicting occupancy and energy usage.	residential	collected data	Presence Schedule

Probabilistic Occupancy Modeling with Markov-Chain Discrete time-steps

This type combines a Markov chain with a Bernoulli process, where each state transition is binary (e.g. active or passive occupant). It excels in binary-state system modeling, like determining whether a room is occupied or unoccupied, where each state is influenced only by its immediate predecessor. Essentially, this specialized Markov model merges a Markov chain with a Bernoulli process, enabling transitions between two distinct states (for instance, active versus inactive occupancy). This model is particularly well-suited for analyzing systems with binary states, e.g. occupancy of a thermal zone (occupied vs. empty), where the current state solely influences the probability of transitioning to the next state. Related to this, studies before 2010 have increasingly focused on the application of probabilistic methodologies, specifically discrete-time Markov chains, to analyze occupant behavior in non-residential buildings. Haldi & Robinson (2011) conducted a pioneering study in Switzerland, concentrating on presence schedules and number of people in the zone. Their approach involved advanced models integrated within the CitySim tool (Thomas et al., 2014), providing insightful data for urban energy planning. Similarly, B. Gunay et al. (2015) explored office spaces in their educational structures in Canada, employing the same methodology to examine not only presence schedules but also lighting demands (i.e. lighting usage schedules) which are dependent on the occupancy in the offices. Their research showed and compared various modeling and simulation approaches on building performance simulation tools. On the other hand, Haldi & Robinson (2009) further developed a similar approach in Switzerland. They tested models including logistic probability distributions to better understand window opening behaviors, expanding the applicability of the occupancy modeling for the building controls. Collectively, these studies underscore the importance of probabilistic models in accurately predicting and managing energy consumption in office environments. However, these models have limitations: they are restricted to binary states and are not suitable for generating activities involving multiple categories of information. Additionally, they exhibit a memoryless property. They assume that the transition probability to the next state depends on the current state which is the proof of disregarding historical data for the next state prediction. Moreover, they are highly sensitive to initial conditions. Nonetheless, these models are not without limitations, as they are constrained to binary states and are unsuitable for generating multi-categorical activity information. Additionally, they exhibit a memoryless property, assuming that the likelihood of transitioning to a subsequent state is dependent solely on the current state and disregards any historical context; they also demonstrate a high sensitivity to initial (i.e. starting) state.

Probabilistic Occupancy Modeling with Markov-Chain First-Order

The first-order Markov chain model operates on the principle that the likelihood of moving to a subsequent state is determined exclusively by the present state, disregarding the influence of any prior states. This approach simplifies the analysis of intricate systems and is particularly valuable when the effects of historical states (i.e. previous states) are diminishing. First-order models are applied as a fundamental tool for the occupancy modeling in the BEM. In the context of office buildings, Erickson et al. (2011) in the USA and Page et al. (2008) in Switzerland have applied the first-order models for presence schedule generation with number of people information. Then, Erickson et al. focused on the USA, where they implemented first-order Markov chains to refine HVAC systems and generate detailed time-series data on occupant presence. Page et al. (2008) also utilized this model to estimate presence schedules, thereby enhancing the accuracy of energy demand predictions**.** Richardson et al. (2008) in the UK advanced this approach by developing a model that produces occupancy time-series data in ten-minute intervals. This level of granularity is crucial for accurately modeling shared energy use e.g. appliance usage in the residential buildings. Wilke et al. (2013) innovatively merged Markov-chain methods with statistical measurements in French households. Their study highlighted the necessity of implementing occupancy models to different occupancy types, thus enabling more precise predictions of time-dependent activities. Collectively, these studies demonstrate the versatility and effectiveness of first-order MC in enhancing our understanding of occupancy patterns and their impact on building energy demand. However, similar to MC with Bernoulli processes, first-order Markov chains also have limitations as these include the memoryless property, where the model's forecasting ability is limited to the current state; these models often show sensitivity to initial states, which can influence the accuracy of their predictions, particularly in scenarios where initial conditions are uncertain or vary significantly.

Probabilistic Occupancy Modeling with Higher-Order Markov-Chain

Higher-order MC are an extension of the first-order model which incorporates a sequence of more than one previous state, to predict future state transitions (Esquivel et al., 2021; Jones & Qin, 2022; van Ravenzwaaij et al., 2018). This added "memory" of the model allows for capturing more intricate relationships within the system, presenting a more dynamic understanding of the sequence, e.g. activities or energy demands. The development of these higher-order models in residential energy studies has brought about significant advancements in the field. Flett & Kelly (2016) from the UK applied this method by introducing a higher-order Markov chain to predict occupancy in the residential buildings. Their research focused on detailed presence schedules and activity patterns, offering a more nuanced understanding of household energy use. In Denmark, Wolf et al. (2019) further expanded the scope of these models. They utilized an inhomogeneous hidden Markov model to simulate zone-level occupancy in residentials. This method provided more detailed modeling for how occupants interact with different spaces within their homes. In Sweden, Widén et al. (2009) applied a non-homogeneous MC to specifically address lighting demand in households thus, their study presents the importance of accurate modeling in reducing unnecessary energy use. Together, their study shows a big step forward in residential energy modeling, highlighting how higher-order Markov chains can offer deeper insights into household energy patterns and occupant behavior. However, with increased sophistication brings challenges. As the model's order rises, the quantity of transition probabilities that need to be estimated escalates dramatically, leading to computational burdens and the risk of overfitting the data (Jones & Qin, 2022), i.e. *order * unique value count(s).* Moreover, the higher-order approach demands extensive data for accurate estimation of these numerous transition probabilities, which may not always be easily obtainable. The interpretability of the model can become compromised, as the impact of past states grows more complex, it becomes increasingly difficult to discern the rationale behind the models predictions, posing a challenge for clear understanding and application of the results.

Probabilistic Occupancy Modeling with Hidden Markov Model

Hidden Markov Models (HMM) assume that the true states of the system are not directly observable; instead, we observe a series of outputs influenced by these states. They are used when the system of interest cannot be directly observed, and the model infers the state sequence through observed data, which is crucial in many real-world applications, including speech recognition and bioinformatics. HMM has emerged as a powerful tool in the field of building energy modeling, with their applications extending across different types of buildings and countries. Liisberg et al. (2016) demonstrated the effectiveness of HMM in the context of Spanish residential buildings. Their approach involved using HMM to classify occupant behavior by analyzing electricity consumption data, providing insights into the patterns of energy use within homes**.** In a different setting, Alfalah et al. (2023) applied HMM to predict occupancy in UK educational buildings. Their work involved training the model under various scenarios, showcasing the adaptability of HMM in understanding occupancy patterns in academic environments**.** Meanwhile, in the residential sector of China, Virote & Neves-silva (2012) leveraged HMM for predicting building energy usage. This study focused on assessing occupant behavior, demonstrating how HMM can be used to accurately forecast energy demands based on how people interact with their living spaces. Collectively, these studies highlight the versatility of Hidden Markov Models in accurately modeling occupant behavior and energy consumption in various building types, contributing significantly to the development of more efficient and sustainable built environments. However, employing HMMs brings certain deficiencies. The computational intensity can be significant for models with numerous hidden states. Additionally, HMMs are based on the assumption that observations within a state are independent, an assumption that may not always hold true in practical scenarios. Similar to the MC models, HMMs are dependent on their initial state, which can impact the accuracy of their predictions. Even if the capabilities are vastly increased for HMMs, they may still face challenges in capturing extensive long-range dependencies among hidden states.

2.5.2.3 Agent-based Modeling (ABM) for Occupancy Modeling

ABM is bottom-up methods for occupancy modeling. This method is suitable for both commercial and residential buildings while including the agents' behavior and interactions. The main advantage of the model is its capability to produce precise schedules and control several behavioral factors simultaneously. It adeptly depicts the autonomous agents' individual and group relationships, offering a nuanced view of occupant behavior. However, it is important to note that this technique is predominantly used in simulation-based models, often facing challenges due to insufficient real data for validation. Moreover, agent-based modeling has limitations for the applicability in large-scale simulations because of the high computational requirements. Various studies have been used as a main model in their studies with different approaches for occupancy modeling.

Table 2.4 Agent-based models of the occupancy modeling

Model	*Approach*	*Building Type*	*Agent type*	*Output*
review paper Li et al. (2019)	ABM models necessitate detailed understanding of specific agent behaviors.	general	Occupants	None
review paper (Grimm et al., 2006)	ODD protocol aims for a universal, uniform ABM documentation framework.	general	Occupants	None
markov chain with bernoulli process (Andrews et al., 2011)	Agent-based simulation models diverse behaviors, focusing on lighting design.	commercial	Occupants	lighting control
markov chain with bernoulli process (Langevin et al., 2014)	Co-simulates occupant comfort, behavior in MATLAB, EnergyPlus for office buildings.	office	Occupants	Presence Schedule (i.e. in office or not)
markov chain with graphical modelling (Liao et al., 2012)	Graphical model depicts occupancy evolution, correlation in multi-zone buildings.	office	Occupants	Presence Schedule (i.e. in office or not)
rule-based (Azar & C. Menassa, 2012)	Simulates commercial buildings energy consumption, accounting for varied occupant behaviors.	commercial	Occupants	Presence Schedule (i.e. in office or not)
rule-based (Lee & Malkawi, 2014)	Simulates individual behaviors, aggregates results for overall building behavior.	commercial	Occupants	Building control
rule-based (Micolier et al., 2019)	Cognitive model simulates user behavior in residential settings.	residential	Occupants	Presence Schedule (i.e. occupancy duration estimation, home or not)

Table 2.4. Agent-based models of the occupancy modeling (continued)

rule-based, Perceptual Control Theory (PCT)	ABM follows standard protocol, compares simulated and real occupant behaviors.	office	Occupants	occupant behavior with builidng control (window, HVAC etc.)
Stochastic (Langevin et al., 2015)	Stochastic models account for daily activities, annual lighting energy levels.	office	Occupants	lighting schedule for building energy simulations
stochastic(Zhou et al., 2015)	Hybrid model integrates agent-based, probabilistic methods for energy waste simulation.	residential	Occupants; Electrical appliances	Presence Schedule (i.e. occupancy duration estimation, home or not)
stochastic(Abdallah et al., 2017)	Generates electricity load profiles for household sustainability potential evaluation.	residential	Occupants; Electrical appliances	Presence Schedule (i.e. occupancy duration estimation, home or not)
Stochastic (Julien Walzberg et al., 2017)	Assesses impact of occupancy patterns on building energy consumption.	residential	Occupants	Presence Schedule (i.e. occupancy duration estimation, home or not)
Stochastic (Chapman et al., 2018)	No-MASS predicts agent presence, activities, and appliance use in homes.	residential	Occupants	Presence Schedule (i.e. occupancy duration estimation, home or not)
population-based, stochastic. (Mosteiro-Romero et al., 2020)	Uses agent-based transportation models for occupancy modeling.	residential	Occupants	Presence Schedule (i.e. occupancy duration estimation, home or not)

ABM stands out for its detailed focus on occupant behavior e.g. window openings, light control etc. Li et al. (2019) provides a review of this field, underscoring the necessity for in-depth understanding of specific behaviors and action sequences of building occupants to develop realistic ABMs. Their work highlights the intricacies involved in simulating the diverse actions of occupants, reflecting the complexity and granularity required in modern building simulations. Grimm et al. (2006) from Germany introduced the Overview, Design concepts, Details (ODD) protocol in their review, aiming to standardize ABM documentation. This universal methodology is a significant stride towards harmonizing agent-based modeling approaches, ensuring consistency and comparability across different studies. Integrating detailed occupant behavior and standardized protocols into building simulations improves accuracy and applicability for energy efficiency.

ABM with Markov-Chain

MCABM are that agents have internal states that can change over time. The probability of an agent transitioning from one state to another depends on the current state. This is governed by predefined transition rules derived from Markov chain principles. These rules represent the likelihood of an agent taking specific actions or behaving in certain ways based on its current state and potentially the state of its environment. Andrews et al. (2011) employ a Markov chain with a Bernoulli process in their ABM simulations. The model is focusing on diverse occupant behaviors and preferences for lighting design for commercial buildings. This approach aims to enhance the usability and efficiency of lighting control systems. Similarly, Langevin et al. (2014) integrated an ABM for occupant comfort and behavior in MATLAB with EnergyPlus for whole building energy simulation, using the *Human and Building Interaction Toolkit* (HABIT). This model is crucial for simulating the adaptive behaviors for the thermal comfort of non-residential building occupants, influencing presence schedules and occupant density. Meanwhile, Liao et al. (2012) propose a graphical model employing a Markov chain to capture occupancy evolution in multi-zone office buildings. Their approach stands out for its lower complexity, yet effectively predicts occupancy patterns. Collectively, these studies showcase the dynamic capabilities of MCABM for simulation of occupant behavior in buildings, ranging from lighting control to thermal comfort and occupancy patterns. However, this approach stands out some deficiencies as for its limited low-order states and reliance on predefined rules.

Rule-based ABM

Building energy simulations use rule-based agent modeling, a bottom-up approach where predefined activity patterns guide virtual occupants interacting with their surroundings. This approach helps understand energy usage patterns in both commercial and residential buildings. Azar & C. Menassa (2012) adopt this methodology to simulate diverse and dynamic energy usage trends among occupants in commercial buildings, involving a comprehensive process from defining behaviors to validating the simulation model. Lee & Malkawi (2014) focus on individual agent behaviors, aggregating them (i.e. group level) to analyze the overall impact on thermal conditions and energy utilization in commercial environments. Their approach highlights the optimization of comfort and energy savings through various occupant behaviors. In the residential domain, Micolier et al. (2019) introduce Li-BIM, a model that simulates the reactive, deliberative, and social behaviors of occupants, thus providing a nuanced understanding of the relation between occupant activities and thermal comfort. Langevin et al. (2015) apply Perceptual Control Theory in their rule-based agent model for office buildings in the USA, aligning simulated agent behaviors with real occupant characteristics to understand the interplay between occupant behavior and building controls. While studies demonstrate the success of rule-based agent modeling in capturing individual behavior and its impact on energy use, limitations exist. This approach often focuses on individual actions and basic appliance interactions, neglecting occupant-to-occupant dynamics.

Stochastic Agent-based Modeling (SABM)

SABM provides a valuable tool for studying complex systems where individual behavior, randomness, and emergent phenomena play a crucial role. While challenges exist in terms of computational cost and data requirements, SABM offers a powerful framework for understanding and predicting the behavior of diverse systems across various fields. Zhou et al. (2015) developed a model to simulate lighting energy use for large office buildings, accounting for the random nature of daily activities of occupants and the lighting power levels. This approach is particularly effective in environments like single offices, where lighting usage patterns are unpredictable. Similarly, Abdallah et al. (2017) introduce a hybrid model that combines agent-based and probabilistic methods to simulate behavioral energy waste at the occupant level in residential settings. This model takes into account variables such as employment type and age, providing a detailed view of energy demand patterns. Julien Walzberg et al. (2017) further contribute to this field by generating household electricity load profiles in smart homes. Their stochastic model focuses on sustainability (i.e. improvement) potential of these homes, emphasizing the importance of lighting schedules and controls in energy simulations.

Y. Chen et al. (2018) implement a stochastic *Occupancy Simulator* to assess the impact of occupancy behavior patterns on energy demand . This model, focusing on residential settings, provides insights into the presence schedules of occupants, highlighting the significant influence of occupant behavior on energy use. Mosteiro-Romero et al. (2020) introduce a novel stochastic population-based approach, utilizing agent-based transportation models for residential occupancy modeling in their study. This approach, named as *PopAp*, is compared with traditional deterministic and stochastic models, highlighting its effectiveness in accurately predicting occupant presence and its impact on energy demand at both annual and lower time scales. This is particularly vital for the planning and operation of district energy systems. Chapman et al. (2018) et al. develop the Nottingham Multi-Agent Stochastic Simulation (No-MASS), employing time-dependent probabilities to generate a synthetic population of agents. This method predicts not only the presence of occupants in residential settings but also their activities, locations, and interactions with building elements like windows, lights, and blinds. By focusing on stochastic elements, these models offer realistic and nuanced simulations of energy demand patterns.

2.5.2.4 Data Driven Modeling

Data-driven approaches have emerged as powerful tools in elucidating the complex patterns of occupancy within buildings as leveraging the vast repositories of data available. Data driven techniques have proven instrumental in uncovering hidden patterns and insightful information critical for creating accurate and reliable models. These methodologies, which thrive on large datasets, have shown a remarkable ability to describe and predict occupant behavior with a high degree of precision. Among the multitude of data mining approaches, decision trees, Bayesian networks, cluster analyses, neural networks have been predominantly utilized for the estimation and identification of occupancy patterns in both commercial and residential settings. The primary advantage of these techniques lies in their ability to investigate recurring trends and logical connections between various variables, with data collection and management processes that are straightforward to execute. Despite these limitations, data driven techniques continue to play a pivotal role for explaining the uncertainty of the occupancy modeling and offer valuable insights for the optimization of building performance and energy efficiency.

Data-Driven Occupancy Modeling, review, and comparison

In the data-driven for occupancy modeling for building energy modeling, significant steps have been made in understanding and predicting occupancy patterns. Z. (Jerry) Yu et al. (2016) provides an overview of data-driven applications by focusing on predictive and descriptive tasks to enhance building performance and energy efficiency. They underscore the potential of data-driven based approaches in analyzing building operational and occupant behavior data.

Table 2.5 Data-driven models of the occupancy modeling

Model	*Approach*	*Building Type*	*Data type*	*Output*
Review (Z. (Jerry) Yu et al., 2016)	Data mining in building engineering focuses on improving energy efficiency	Various	Building operational data	None
Comparison (Huchuk et al., 2019)	Machine learning models predict occupancy using connected thermostat data	residential	multiple PIR motion sensors data.	future occupancy states in homes
Comparison (B. Yang et al., 2021)	Framework predicts seasonal occupancy in residences using machine learning	residential	Measurement data	Presence Schedule
Clustering (Z. Yu et al., 2011)	Methodology uses cluster analysis to link behavior with energy use	residential	Measurement data	Presence Schedule
Clustering (Buttitta et al., 2019)	Novel archetypes integrate varying occupancy schedules within each model	residential	Time-Use-Survey	Presence Schedule
Clustering (Liang et al., 2016)	Office occupancy schedules predicted using data mining and clustering	office	Measurement data	Presence Schedule
clustering, hierarchical (Aerts et al., 2014)	Study identifies seven distinct occupancy patterns in residential settings	residential	Time-Use-Survey	Presence Schedule

Table 2.5 Data-driven models of the occupancy modeling (continued)

clustering and Association rule mining (ARM) (Ashouri et al., 2018)	Advisory system uses data mining for energy-saving recommendations	residential	Measurement data	Presence Schedule
clustering, k-Shape, Support Vector Regression (SVR)	K-shape clustering improves energy consumption forecasting accuracy	institutional	Measurement data	Presence Schedule
clustering, k-Shape (J. Yang et al., 2017)	Framework extracts occupant schedules for residential energy simulation	residential	Measurement data	Presence Schedule
classification, decision tree (Panchabikesan et al., 2021b)	Data mining method learns occupant behavior from office appliance data	office	Office appliance power consumption data	Presence Schedule
classification, decision tree (Zhao et al., 2014)	Framework translates office occupancy patterns into energy modeling profiles	office	Measurement data	Presence Schedule
classification, decision tree (D"Oca & Hong, 2015)	Data mining methods explore occupant behavior in affordable housing	residential	Measurement data	Presence Schedule
neural network (Ren et al., 2015)	Framework predicts institutional energy consumption using machine learning	institutional	Measurement data	Presence Schedule
neural network (Migliori et al., 2022)	ANN trained with simulated data predicts post-occupancy energy consumption	institutional	Measurement data	Presence Schedule
neural network, Generative Adversarial Network (Z. Chen & Jiang, 2018)	GAN framework innovates in building occupancy modeling	office	Measurement data	Presence Schedule
neural network, spiking (SNN) (M. Wang et al., 2015)	SNN model detects occupancy in green building automation systems	office	Measurement data	Presence Schedule

Building upon this, Huchuk et al. (2019) compare various machine learning models, including random forests and recurrent neural networks, for predicting occupancy states in residential buildings using multiple PIR motion sensor data, which is one of the contemporary approaches recently and it has a potential to build a foundation for Internet-of-Things (IOT) applications. Their study, while insightful, is limited by the number of occupants detected. Further, B. Yang et al. (2021) develop a season-customized occupancy prediction framework using machine learning algorithms and recursive feature elimination. Their approach, tested in a single residential unit, demonstrates over 85% accuracy in occupancy prediction across different seasons, though the results look promising the accuracy levels are lower than the general confidence interval levels, i.e. 5% error rate.

Data-Driven Occupancy Modeling with Clustering

In the exploration of clustering techniques for occupancy modeling, Z. (Jerry) Yu et al. (2016) developed a methodology using cluster analysis to investigate the impact of occupant behavior on residential energy demand. Their focus on occupancy duration estimation (i.e. presence schedule), however it was limited in quantifying the number of people. Similarly, Buttitta et al. (2019) worked on defining occupancy-integrated archetypes for residential settings, integrating various occupancy schedules within each archetype. This is a genuine approach by considering the uncertainty comes from occupancy for the archetype generation. On the other hand, their innovative use of TUS data provided a deeper understanding of occupancy patterns, though it was similarly constrained to only estimating the number of occupants besides presence schedules. Liang et al. (2016) implemented clustering to predict occupancy schedules in office buildings. Despite challenges with data reliability due to sensor issues, their approach enhanced the accuracy of occupancy predictions.

D. Aerts et al. identified seven distinct occupancy patterns in residential buildings using hierarchical clustering, offering a detailed view of presence schedule patterns derived from Time-Use-Survey data. A similar approach has been used in this thesis to compare with the proposed sequential neural network model and hierarchical clustering approach. In a different approach, Ashouri et al. (2018) combined clustering with Association Rule Mining (ARM) for energy demand analysis in residential buildings, which involves identifying patterns of energy consumption. Their advisory system showcased potential energy savings by influencing occupant energy usage, achieving actual savings of 12% in their case study. Focusing on institutional buildings, J. Yang et al. (2017) applied a combination of k-shape clustering, which groups similar time-series data points together, and Support Vector Regression (SVR), a machine learning technique for prediction, to identify energy demand patterns. This approach substantially improved forecasting accuracy in energy models for the majority of the buildings studied. Finally, Panchabikesan et al. (2021b) proposed a framework to extract varying occupant schedules in residential buildings, demonstrating the effectiveness of lighting and plug load data in representing occupant activity.

In conclusion, while clustering models offer a simple and efficient way to group similar data points, their limitations in handling complex relationships, explaining cluster formation, and dealing with high-dimensional data can be overcome by utilizing advanced neural networks.

Data-Driven Occupancy Modeling with Classification

Several research studies have significantly contributed to the development of classification techniques used in occupancy modeling. Zhao et al. (2014) utilized a decision tree model to learn occupant behavior in offices from appliance power demand data (i.e. collected data). They have achieved high accuracy in occupant behavior classification and grouping the occupancy schedule in terms of correlations. However, their work was constrained by the reliance on occupant behavioral consistency which cannot widely capture the changes in the occupant behaviors. D'Oca & Hong (2015) also applied a decision tree approach, focusing on translating occupancy patterns in offices into working individual profile schedules for energy modeling. Their method provided valuable archetypal profiles, though its applicability might be limited to the specific dataset studied as it cannot be generalized. In a residential context, Ren et al. (2015) explored occupant behavior and heating system operations in affordable housing using similar data-driven techniques. They successfully identified room temperature patterns linked to heating energy demand, yet their findings were specific to the affordable housing dataset. Each study, while advancing the understanding of occupancy patterns in different building environments, highlights the challenges in ensuring data reliability and generalizability of results. Although these studies offer valuable insights into occupancy patterns, decision tree models used in their research have limitations. While they provide a clear and understandable approach, they struggle to capture intricate relationships, adapt to changing behaviors, and apply their findings broadly. Advanced neural networks can overcome these limitations, offering a more powerful and adaptable approach for occupancy modeling.

Data-Driven Occupancy Modeling with Neural Networks

Data-driven modeling represents a paradigm shift in occupancy studies as putting the real-time datasets (i.e. either collected or measured) into the center of the modeling process. The process involves using a variety of data sources to improve the accuracy of occupancy predictions (Tuniki et al., 2021) e.g. sensors, population statistics etc. This chapter presents the importance of data-driven modeling in the context of residential buildings by exploring how the availability of high-resolution data enables to gain deeper insights into occupant behavior and its impact on energy demand and thermal comfort (Gaetani et al., 2016).

In the field of data-driven occupancy modeling, neural networks have increasingly become popular for their advanced capabilities for classification and prediction. Anand et al. (2021) and Z. Chen & Jiang (2018) have significantly advanced occupancy modeling using neural networks in different building environments. Anand's team developed a machine learning and neural network-based framework for predicting energy demand for institutional buildings. They have focused on occupancy counting errors for number of people information and improving hourly electrical energy demand predictions. However, their model may require modifications for different building types or locations as it is only applied for one building type. Z. Chen & Jiang (2018), on the other hand, introduced a Generative Adversarial Network (GAN) for occupancy modeling in office buildings. Their GAN framework successfully outperformed traditional models i.e. deterministic and probabilistic models.

Similar to other studies Migliori et al. (2022), and M. Wang et al. (2015) each utilized neural networks to tackle different aspects of occupancy detection and energy demand prediction. Migliori's team used an artificial neural network (ANN) trained with simulated data to predict energy demand in institutional buildings, noting a deviation from actual post-construction energy consumption in their initial model. Wang's team proposed a Spiking Neural Networks (SNN) model for occupancy detection in office building automation systems, although their model's accuracy has yet to be validated with real sensor data.

2.5.2.5 Limitations of Current Data-Driven Occupancy Modeling Approaches

Data-driven occupancy modeling has its challenges and limitations. A major challenge is the privacy and security concerns associated with collecting detailed occupancy data (Akkaya et al., 2015). Scalability of these models can be challenging due to the cost and resources required for extensive data collection (S. Chen et al., 2021). For instance, it is easier to apply data collection for a single building experiment compared to the urban level. Adapting to rapidly changing occupant behavior in dynamic environments can also be challenging (Gaetani et al., 2016). In addition, the diversity of occupant behaviors and building types can complicate the creation of universally applicable models, often requiring customization (H. B. Gunay et al., 2013). For example, there are many different occupant profiles that can exist simultaneously in a thermal zone. Despite these challenges, ongoing research aims to address these limitations and improve the practicality and effectiveness of data-driven modeling.

Despite the advancements in neural network-based occupancy modeling, a notable limitation across these studies is the lack of a comprehensive evaluation of complete occupancy modeling. Existing models fail to consider the full picture, neglecting factors like occupant demographics, schedules (both presence and activity), and energy-related schedules. This gap highlights the need for more holistic approaches in occupancy modeling. Such integrative models would not only enhance the accuracy of predictions but also provide a more nuanced understanding of the interplay between various factors influencing energy demand in buildings. The development of these comprehensive models remains a crucial area for future research, aiming to capture the full spectrum of occupancy dynamics.

2.5.3 Literature Gaps in the Current Occupancy Modeling Approaches

Previous chapters have presented the limitations inherent in existing occupancy modeling methodologies, e.g. deterministic, statistical, agent-based, and probabilistic. A critical deficiency lies in the lack of detailed modeling by capturing every action of the occupants and their residential status. To elaborate further, this involves addressing long-term dependencies and comprehensive modeling that considers schedules, demographics, and activities concurrently.

Current models often rely on excessive assumptions, leading to a multi-step process that accumulates errors at each stage. For instance, the sequential use of clustering followed by MC models introduces compounded inaccuracies, ultimately diminishing the accuracy of the final output. This fragmentation underscores the necessity for a more integrated approach. A comprehensive and realistic evaluation of occupancy demands a holistic methodology that encompasses demographics, individual-level modeling, and considers the intricate interactions among these factors.

To address these gaps, this thesis proposes a deep learning model utilizing Long Short-Term Memory (LSTM) networks. This model is developed to classify daily activity schedules and locations of occupants in urban residential buildings. Operating on a dataset with every hour per day, representing distinct time intervals, it aims to perform multi-categorical classification for each interval. This method not only promises enhanced accuracy in occupancy modeling but also provides a nuanced understanding of occupant behavior in urban residential contexts.

2.6 How to Realistically Model the Occupancy

Occupancy modeling has been limited by methodologies that focus on simplicity over accuracy. The common practice of using linear algorithms to predict residential occupancy patterns exemplifies this trend. Although simple, these approaches do not account for the inherently unpredictable nature of human occupancy. This oversimplification leads to a significant gap in modeling accuracy because it ignores the complex, variable patterns of human presence and activity within buildings (S. Chen et al., 2021). The need for advanced modeling techniques that reflect the dynamic and unpredictable nature of occupancy has become increasingly apparent.

A conventional error about the occupancy modeling is the isolated modeling of different components of occupancy, e.g. presence and activity. This segmented approach neglects the dependent nature of occupant behavior. Advanced modeling techniques, capable of capturing the detailed relationships between different aspects of occupancy, are essential. Methods that effectively predict or classify sequential data can offer a more holistic and accurate representation of occupant behavior. Such integrated modeling is crucial for understanding the complex dynamics of occupancy and its impact on energy demand.

The intricate relationship between occupancy and energy demand necessitates a detailed and high-resolution analysis. Occupants act as active agents within residential settings, influencing energy demand through a variety of activities, ranging from heating to lighting (Tuniki et al., 2021). <u>It is necessary to analyze occupant activities and corresponding energy demands at an hourly or even higher resolution to accurately model this relationship.</u> This approach acknowledges the energy usage in residential buildings, driven by the patterns or hourly changes of occupancy. A refined understanding of this relationship is key to enhancing the accuracy and effectiveness of UBEM.

An overlooked aspect in occupancy modeling is the dynamic interaction between occupants and their engagement with residential buildings and appliances

(Heydarian et al., 2020). Occupants are not solitary entities; their activities are often interrelated and affect each other in different ways. For instance, the use of lighting fixtures, appliances, and heating or cooling systems is not just a matter of individual preference but is also shaped by the presence and activities of other occupants. Additionally, features like windows and thermostats are not passive elements; they are dependent on occupant's daily routines and comfort, and their usage patterns offer valuable insights into energy demand. Advanced occupancy modeling should encapsulate these multifaceted interactions and dependencies (Heydarian et al., 2020). Recognizing and accurately modeling these complex interplays is essential for a more complete understanding of energy demand within residential buildings. This approach not only enhances the accuracy of UBEM but also provides a deeper insight into the potential for energy-saving strategies tailored to the rhythms and habits of building occupants.

2.7 Data Collection for the Occupancy Modeling

Accurate occupancy modeling relies heavily on systematic data collection (Akkaya et al., 2015). A variety of techniques are used in this process (Yan et al., 2017), ranging from surveys and questionnaires that gather information directly from the occupants themselves, to the integration of sensor technologies, smart meters, and Internet of Things (IoT) devices that enable real-time capture of occupancy patterns. The chapter also explores the incorporation of social and environmental data, highlighting the impact of external factors on occupant behavior. It also examines data management practices, including preprocessing and integration, which are critical to improving data quality and reliability. This chapter also pays special attention to occupant consent and privacy considerations, emphasizing the ethical principles that guide responsible data collection. By navigating the intricacies of data collection, researchers and practitioners are equipped with indispensable tools for building accurate and actionable occupancy models.

2.7.1 Data Quality for the Occupancy Modeling

Urban level building occupancy have a relation with the content and the quality of the datasets. Due to the lack of data availability, estimating the building occupancy profile at the urban scale is challenging (Dong et al., 2022; T. Hong et al., 2020). Thus, most of the previous research has conducted BEM with fixed building occupant profiles. This can cause false calculations for BEM.

Although numerous techniques and datasets exist for estimating building occupancy profiles, challenges remain when estimating occupancy on an urban scale. As previously noted, building occupancy can fluctuate due to various factors, e.g., building use type, age, size, location, day of the week, and season. Recently, innovative data sources have provided new opportunities for a more comprehensive understanding of occupant patterns in buildings.

The occupancy domain is defined by the type of datasets used for extracting occupancy information, which can be classified under two categories, i.e., 'explicit' and 'implicit'. Explicit datasets allow for direct counting of occupants. Implicit datasets serve as proxies for occupant numbers by using environmental indicators e.g., as temperature, humidity, pressure, and carbon dioxide levels.

2.7.2 Preprocess for Occupancy Modeling

Efficient data handling plays a crucial role in the process of gathering data, guaranteeing the reliability and precision of the data utilized for occupancy modeling (Dabirian et al., 2021; Happle et al., 2018; Tuniki et al., 2021). Data preprocessing is an essential first step that involves a meticulous approach to cleaning, filtering, and refining raw data. The goal of this process is to improve data quality by eliminating inconsistencies, inaccuracies, and noise that can affect the accuracy of occupancy models. Proper preprocessing not only improves data quality, but also streamlines subsequent analysis and modeling tasks (Dabirian et al., 2022). On the other hand, data integration is another fundamental facet of data management. It involves the seamless merging of information from different sources, combining data sets into a unified structure. The integration process orchestrates disparate pieces of data to create a holistic and comprehensive view of occupancy behavior. This synthesized view is valuable for the modeling process due to providing a comprehensive perspective that encompasses the many aspects of how occupants interact within residential spaces (Heydarian et al., 2020; Micolier et al., 2019). Effective data management practices, including robust preprocessing and integration, are essential pillars that maintain data integrity and support the accuracy of occupancy models.

2.7.3 Data Collection process for Implicit Dataset

Babaei et al. (2015) provided a comprehensive overview of energy demand data collection, categorizing datasets sourced through measurement, survey, and simulation methods. They created a directory of these resources, though they noted challenges in standardizing data reporting. Bianchi et al. (2020) introduced "Parametric Schedules" for modeling occupancy-driven schedules in commercial buildings using metered electric demand data. They also stated about limitations for buildings with residential load profiles. Zhuang et al. (2022) developed an autoencoder Bayesian LSTM model for probabilistic occupancy prediction in educational buildings, integrating uncertainties for energy-efficient ventilation management based on the reliance on CO_2-based estimations. Ekwevugbe et al. (2013) achieved a 75% occupancy estimation accuracy in office buildings using a novel sensor network and neural network analysis. They also highlighted the need for improved accuracy during occupied periods. Lastly, B. Yang et al. (2021) employed data mining and machine learning techniques for high-accuracy occupancy prediction in residential settings, with the caveat of limited testing environments and marginal accuracy improvements over consecutive prediction models

Table 2.6 Implicit Datasets for Occupancy Modeling

dataset type	*Data Collection*	*Approach*	*Input Data type*	*Output*
implicit dataset (Babaei et al., 2015)	Electricity Demand	Overview of building energy data: measurement, survey, simulation	Energy data: measurement, survey, simulation methods.	Energy consumption data set directory.
implicit dataset (Bianchi et al., 2020)	Electricity Demand	Parametric Schedules model occupancy in diverse buildings efficiently	Metered electric data infers occupancy schedules.	Energy consumption data set directory.
implicit dataset (Zhuang et al., 2022)	Environmental data	Autoencoder LSTM model predicts occupancy with uncertainty integration	Educational building data: CO2, temperature, lighting, fans.	Probabilistic forecasts optimize ventilation, reduce risk.
implicit dataset (Ekwevugbe et al., 2013)	Environmental data	Innovative sensor network uses environmental data for occupancy	Data from sound, CO2, temperature, motion sensors.	Real-time open-plan occupancy estimation.
implicit dataset (B. Yang et al., 2021)	Environmental data	Data mining framework predicts occupancy, adjusts seasonally	Indoor/outdoor data: environment, movement, energy, time.	Seasonal occupancy predictions, varying accuracy.
implicit dataset (Sheikh Khan et al., 2021)	Passive Infrared sensors	PIR sensors under desks track office occupancy	PIR sensor data shows occupancy, count.	Occupancy schedules inform HVAC design.
implicit dataset (Raykov et al., 2016)	Passive Infrared sensors	Single PIR sensor estimates occupancy with machine learning	Occupancy data from single PIR sensor.	Room occupancy updated every 30s.
implicit dataset (Huchuk et al., 2019)	Building dataset	Machine learning predicts home occupancy from thermostat data	Thermostat data: motion, time, weekday/weekend.	Predictive home occupancy for HVAC control.
implicit dataset (Burak Gunay et al., 2015)	Building dataset	Occupancy-learning algorithm adjusts temperature based on motion data	Motion sensors track office occupancy patterns.	Adaptive temperature schedules for energy optimization.
implicit dataset (Picaut et al., 2019)	Building dataset	Crowdsourced app collects sound data for noise mapping	Acoustic, GPS, perceptual data from NoiseCapture app.	Community noise maps on web platform.

Sheikh Khan et al. (2021) utilized a PIR sensor under desks to accurately detect occupant presence and count in office spaces, achieving 87.5% accuracy but overestimating occupant count. This method provides valuable insights for HVAC sizing and office space utilization, despite needing refinement in sensor placement. Raykov et al. (2016) used a single PIR sensor and nonparametric machine learning algorithms for room occupancy estimation in office settings, achieving occupancy estimates within ±1 individual but facing challenges with sensor limitations. Huchuk et al. (2019) evaluated machine learning models (i.e. random forest), using thermostat interval data from homes to predict occupancy states. They also stated the need for further research with ground-truth data. Burak Gunay et al. (2015) developed a recursive occupancy-learning control algorithm for temperature setback schedules in office and institutional buildings. They achieved 10-15% reductions in annual heating and cooling loads, though its applicability to larger scales remains unexplored. Lastly, Picaut et al. (2019) introduced a novel crowdsourcing method for noise mapping using smartphone data in city landscapes. They showcased the data consistency and quality across various devices in the similar setups.

2.7.4 Data Collection Process for Explicit Dataset

Advanced sensor and camera technologies is at the core of real-time occupancy-related data collection, e.g. occupancy sensors, motion detectors, and environmental sensors (Akkaya et al., 2015). These sensors are equipped with smart meters, which provide detailed information on energy consumption and help correlate energy use with occupancy patterns. Internet of Things (IoT) devices, such as smart thermostats, lighting controls, and home automation systems, provide valuable insights into occupant interactions with building systems (Nejadshamsi et al., 2023). Together, these technologies provide a wealth of data for building accurate occupancy models.

Tomastik et al. (2008) developed a non-linear stochastic state-space model using video inputs from cameras. They integrated video signal processing and people traffic models for building occupancy estimation in offices. They also presented the limitations about varying camera qualities, e.g. resolution. Petersen et al. (2016) achieved over 98% accuracy in real-time occupancy detection using infrared depth frame images from a Kinect camera in office environments. However, there were also potential misidentifications for occupant detection. Pang et al. (2018) proposed a mobile-internet-based occupancy detection method using location-based services and GPS data for building energy simulation in offices. Parker et al. (2017) utilized personal location metadata from Google to create occupancy schedules for commercial facilities. Happle et al. (2018) compared context-specific building occupancy schedules derived from location-based services data to standard schedules, revealing overestimations in standard schedules in commercial settings.

Y. Wang & Shao (2017) implemented a Wi-Fi based indoor positioning system in an institutional building, identifying significant occupancy pattern changes with potential energy savings.

Table 2.7 Explicit Datasets for Occupancy Modeling, others

dataset type	*Data Collection Type*	*Approach*	*Input Data type*	*Output*
explicit dataset Tomastik et al. (2008)	cameras	Non-linear model estimates occupancy with extended Kalman filter	Video feeds count people per zone.	Video feeds count people per zone.
explicit dataset Petersen et al. (2016)	cameras	Kinect camera for unsupervised, privacy-focused occupancy detection	Kinect camera captures room depthframe images.	Anonymous room entry and exit counts.
explicit dataset Pang et al. (2018)	location-based service, GPS	Mobile-internet method detects occupancy for energy simulation	Mobile-internet occupancy, historical electricity data from BAS.	Improved energy simulation with building model.
explicit dataset Parker et al. (2017)	location-based service, GPS	Uses metadata for non-domestic occupancy scheduling	Google data provides facility popular visit times.	Custom occupancy schedules enhance model accuracy.
explicit dataset (Happle et al., 2020)	location-based service, Cellular	LBS data creates context-specific occupancy schedules	Location-based data from U.S. retail, restaurants.	Context-specific schedules for energy simulations.
explicit dataset (Y. Wang & Shao, 2017)	location-based service, Cellular	Wi-Fi system monitors university library occupancy patterns	Wi-Fi, lighting sensor data indicate occupancy.	Occupancy profiles boost efficiency, productivity.
explicit dataset (Shao et al., 2017)	location-based service, bluetooth	RFDoorGuard identifies individuals with Bluetooth beacons, RSSI signals	Door access monitored with Bluetooth beacons.	Device-free individual room access identification.
explicit dataset (Park et al., 2018)	location-based service, bluetooth	Open-source Bluetooth system detects occupancy with Raspberry Pi	Bluetooth signals detect occupant presence.	Building occupancy estimation, identifies occupant types.

The applicability of study is limited to one room type and season. Shao et al. (2017) introduced RFDoorGuard, a Bluetooth-based door access system in office settings, with accuracy decreasing in larger groups. Lastly, Park et al. (2018) proposed a low-cost Bluetooth occupancy detection system using a Raspberry Pi setup in institutional buildings. They also showcased the limitations in motion direction detection and privacy concerns.

Richardson et al. (2008) generated detailed ten-minute interval occupancy data for residential settings, ie., distinguishing weekdays from weekends as a fundamental property of TUS datasets. However, the model doesn’t account for day-to-day consistency in occupant behavior. Flett & Kelly (2016) developed an occupancy model by type, enhancing accuracy by considering occupant interactions and using a higher-order Markov Chain process. Wolf et al. (2019) applied an inhomogeneous hidden Markov model for room-level occupancy simulation by reflecting Danish behavioral patterns, but with limitations of ignoring the inter-occupant interactions. Wilke et al. (2013) modeled time-dependent activity probabilities based on French TUS data. They aimed to improve productivity by including individual characteristics and transitions between activities. Aerts et al. (2014) identified seven distinct occupancy patterns in residential settings using Belgian survey data for building energy simulations. The study has some deficiencies in accounting for inter-household dynamics. Mitra et al. (2020) used ATUS data to establish typical U.S. occupancy patterns, highlighting significant differences from current energy modeling methods, though the combination of ATUS data might affect accuracy.

Table 2.8 Explicit Datasets for Occupancy Modeling, Questionnaire

dataset type	*Data Collection Type*	*Approach*	*Input Data type*	*Output*
explicit dataset (Richardson et al., 2008)	Questionnaire survey	Model generates ten-minute interval occupancy data	Time-Use-Survey	Occupancy duration estimation, inside or not.
explicit dataset (Flett & Kelly, 2016)	Questionnaire survey	Occupancy model differentiates by occupant, household types	Time-Use-Survey	Occupancy duration estimation, inside or not.
explicit dataset (Wolf et al., 2019)	Questionnaire survey	Applies inhomogeneous Markov model for occupancy simulation	Time-Use-Survey	Occupancy duration estimation, inside or not.
explicit dataset (Wilke et al., 2013)	Questionnaire survey	Models time-dependent home activity probabilities	Time-Use-Survey	Occupancy duration estimation, inside or not.
explicit dataset (Aerts et al., 2014)	Questionnaire survey	Study finds seven residential occupancy patterns	Time-Use-Survey	Occupancy duration estimation, inside or not.
explicit dataset (Mitra et al., 2020)	Questionnaire survey	ATUS data establishes typical U.S. home occupancy patterns	Time-Use-Survey	Occupancy duration estimation, inside or not.

While camera and location-based service methods offer remarkable accuracy and potential for development in occupancy modeling, they encounter notable privacy concerns. These methods, although technologically advanced, must navigate the delicate balance between data collection and individual privacy rights. On the other hand, survey-based studies provide a more comprehensive and realistic dataset that can be tailored to specific national contexts. These surveys allow for a deeper understanding of occupancy patterns and anomalies within the building stock, offering insights that are less intrusive and more culturally sensitive. This confusion highlights the ongoing challenge in occupancy modeling: integrating precise, innovative data collection techniques while respecting privacy and adapting to diverse cultural environments.

2.8 Occupancy-Related Features

Occupancy features include whether people are present, their appliance and lighting usage, preferred temperature settings, hot water needs, and how they open and close windows (Happle et al., 2018). The following sections explore how each of these occupant behaviors affects energy use at both individual building and city-wide levels.

The presence of occupants was taken into account to determine the associated internal heat gain (K. Ahmed et al., 2017). As previously discussed, plug load, lighting load, setpoint schedules, and Domestic Hot Water (DHW) usage are additional occupant-related inputs directly linked to energy types. It is crucial to assign appliance and lighting load schedules based on occupancy, as they are closely correlated. These inputs depend heavily on the presence of occupants to calculate internal heat gains.

A strong correlation exists between appliance and lighting load schedules and occupancy in residential buildings (Liang et al., 2016). The plug load schedules, indicative of electrical equipment usage, are closely tied to the occupancy schedule, a relationship clearly demonstrated in high-performance mixed-use buildings (Bottaccioli et al., 2019; Mosteiro-Romero et al., 2020). In terms of lighting schedules, which significantly influence energy demand, the interaction between occupants and lighting systems correlates with lighting energy use. Studies have shown that lighting load variations mirror occupant activity patterns, varying across buildings (Zhou et al., 2015). Richardson et al. (2009) developed a high-resolution lighting demand model, incorporating natural lighting and occupant activities, to create detailed electricity demand profiles for residential buildings based on active occupancy schedules. Recently, researchers have begun exploring the use of equipment energy data as a data-driven approach to improve occupancy modeling (Bottaccioli et al., 2019).

Temperature setpoint schedules are mostly modeled for heating set point/back and cooling set point/back. They are vital in simulating HVAC system performance, influenced by environmental, building systems, and occupant-related factors (Wei et al., 2014). Studies like those by Booten et al. (2017) indicate that minor variations in setpoint temperatures can significantly impact energy demand predictions. Domestic Hot Water (DHW) usage is another key factor, and it is influenced by climatic conditions, socio-economic factors, and building types. Techniques for generating DHW profiles range from stochastic models to advanced data-driven models (Bakker et al., 2008), i.e. ANN. Window operations are impacting indoor air quality, room temperature due to ventilation exchanges.

A window actions (i.e., opening or closing) model considering interaction probability and consequences was developed by Robinson et al. (2007). Its output is integrated into the whole building thermal model, leading to deterministically calculated ventilation rates based on window opening schedules.

Finally, building energy simulations capture the complex interplay between occupant behavior and energy use by incorporating factors like presence, appliance and lighting usage, preferred temperatures, hot water needs, and window operation. This study specifically considers occupant presence, activity, lighting, and equipment schedules as key inputs to accurately predict energy demand in urban buildings.

2.9 Occupancy Integration for UBEM

UBEM tools incorporate several occupancy-related parameters, each of which plays a critical role in accurately simulating energy demand in residential environments. Parameters such as internal gains, controls, densities, and characteristics are the fundamental elements that influence the energy modeling process (El Kontar & Rakha, 2018). Together, they account for the thermal and energy impact of occupants in a building. Detailed household and occupant profiles provide a high resolution of the living patterns within residential units, e.g. the number of occupants, their typical schedules, and lifestyle characteristics (Heydarian et al., 2020; Schweiker et al., 2017). The inclusion of specific schedules for presence, activity, and energy provisioning allows for a temporal dimension in modeling, capturing the dynamic nature of occupancy and its direct impact on energy demand.

HVAC, heating, cooling, lighting, and equipment schedules are integrated to reflect the actual usage patterns of various systems within a residential setting (Tuniki et al., 2021). These schedules are important in modeling the operational times and intensity of energy-demand systems, closely aligning with the occupants' daily routines. Preferences play a significant role in this context, encompassing set points and setback times for heating and cooling systems, illuminance levels for lighting, and thermal comfort parameters. These elements are not just about energy demand; they also represent the comfort and lifestyle choices of the occupants. The accurate modeling of these preferences is key to understanding how energy demand is shaped not only by the physical presence of occupants but also by their comfort requirements and behavioral patterns. This holistic approach in incorporating various occupancy-related features into UBEM tools enhances the precision and relevance of energy demand predictions, enabling more efficient and occupant-centric energy management in urban residential buildings.

CHAPTER 3

METHODOLOGY

In this chapter, a novel methodology is presented for generating accurate occupancy schedules for residential buildings to eliminate the *performance gap* which is discrepancy between actual and simulated energy use. Thus, a deep learning model methodology has been used for advanced energy modeling including sequential data classification using LSTM networks, designed to classify the daily activities of urban residents, annually. Using the CENSUS and TUS datasets from the Italian National Institute of Statistics (ISTAT), the research intricately combines temporal and non-temporal data to reflect the diverse behaviors of residents.

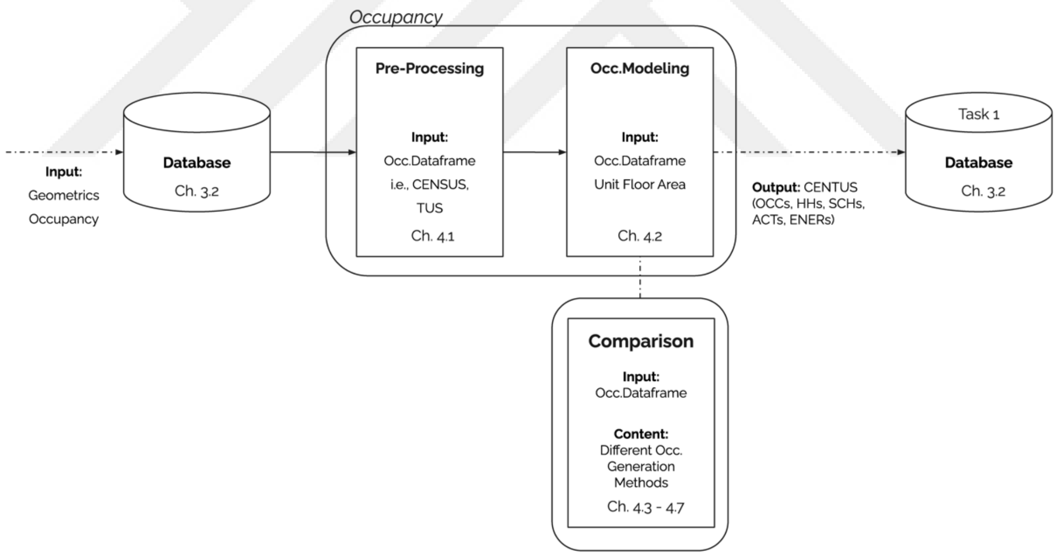

Figure 3.1. Flowchart of the Thesis Study

Figure 3.1 is explaining the structured methodology of the thesis, beginning with the aggregation of building data in Task 1 and transitioning to Task 2, where spatial attributes are also pre-processed into a DataFrame. Concurrently, Task 3 refines occupancy data into an "Occupancy DataFrame" for use in Task 4, the central occupancy modeling stage, which uses floor area and occupant profiles to predict residential energy demand. The simulation in Task 5 incorporates this occupancy data to simulate residential unit energy demand, culminating in Task 6 where the simulated occupancy-induced energy demand is compared to standard benchmarks, i.e. generated occupancy vs. default occupancy. The entire process, from data collection to comparative analysis, underscores an approach to modeling that is essential to improving the accuracy of urban energy demand forecasts and closing the "performance gap" in current UBEM process.

3.1 Research Methodology

This work presents a novel methodology that significantly develops accurate occupancy modeling in the context of urban residential building stock. At the center of this approach is the development of occupancy data as conversion with preprocessing and augmentation from the of population statistics to the occupancy modeling input.

3.1.1 Development of Dataset for Occupancy Modeling

To generate detailed occupancy models, a comprehensive dataset is needed by emphasizing the integral role of various types of data in UBEM. This dataset composes of building characteristics, meteorological elements, and energy use patterns besides occupancy-related features, see in Chapter 2.8. The flowchart in Figure 3.1 illustrates the systematic process of data transformation and integration that leads to the creation of a robust occupancy model. This includes harmonizing data sets from disparate sources and calibrating simulation inputs to reflect the unique conditions of each dwelling unit.

Central to the thesis is the use of the TUS temp dataset, derived from the ISTAT, which provides a mixture of temporal and non-temporal data elements (ISTAT, 2013a, 2018, 2022). This dataset is particularly valuable for its granularity and demographic diversity, which are critical to developing a nuanced and scalable occupancy model. The unique structure of this dataset, which includes both household and occupant identifiers, allows for an in-depth analysis of occupant activities in relation to energy demand.

3.1.2 Sequential Occupancy Information Modeling

The methodology extends traditional data models by augmenting the dataset with different activity schedules across seasons and days, providing a more comprehensive and realistic representation of daily routines. This seasonal and daily variability is critical to capturing the fluctuating nature of energy demand and is a distinguishing feature of this thesis.

This model is designed to classify the daily activity schedules of urban residential occupants, processing data in 30-minute intervals per day (i.e. 48 rows of set data) to reflect the daily occupant activity. This temporal resolution is critical in capturing the dynamics of occupant behavior throughout the day and is a significant advancement over traditional hourly or daily averaging models.

3.1.3 Occupancy Integration to Building Energy Modeling

The thesis outlines a detailed process of urban building energy simulation using Python, starting with the transformation of .osm (OpenStreetMap) data into .idf (Input Data File, EnergyPlus) formats, and culminating in complex energy simulations for residential buildings with multiple thermal zones (Crawley et al., 2000; LBNL, 2009). The integration and the simulations are executed with the help of advanced tools and libraries, i.e. geopandas, shapely, osmium, eppy, and geomeppy. (Philip et al., 2011a). These tools provided not only increases the precision of data manipulation, but also ensures the seamless integration of geometric, construction, and scheduling parameters into the energy modeling process.

3.1.4 Analyzing the Results and Outlining Next Steps

The final section discusses the organization of this integrated data into a structured, easily navigable, and efficient database, illustrating its critical function in UBEM applications.

3.2 Occupancy-Related Datasets

The study used Italian population statistics datasets derived from STAT, i.e. the 15th Population and Housing CENSUS 2011 and the "Use of Time" survey 2013-2014 (ISTAT, 2013a, 2018). The CENSUS dataset provides detailed information about Italy's population, family conditions and arrangements, residential properties. The "Use of Time" survey is a Multipurpose Household Surveys. This presents in-depth insights into how individuals (i.e. occupants) allocate their time across various activities during daily and weekly schedules.

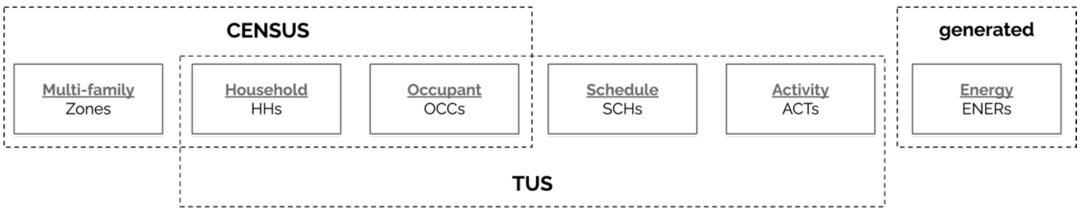

Figure 3.2. Occupancy-related Dataset, Content

Here is the content of the occupancy datasets as shown in Figure 3.2. CENSUS datasets are used for selecting the household formations form the preprocessed TUS datasets according to residential properties. TUS datasets provided a comprehensive view of the Italian residential landscape for accurate modeling of occupancy and energy usage in residential buildings.

3.2.1 The Content of the Dataset, HETUS and Content of TUS

The Harmonized European Time Use Surveys (HETUS) guidelines are initiated by Eurostat in 2000 and then it is updated in 2008 and 2018. Eurostat have standardized TUS across Europe (Eurostat, 2019). These guidelines have evolved:

- to simplify data collection processes
- to enhance compatibility of concepts

In 2018, they have updated of the streamlined survey methodologies and set the stage for future integration of advanced data collection tools, e.g. web diaries and mobile apps. This can facilitate the data collection process in terms of accuracy, time, and scalability. This continuous evolution of HETUS reflects a commitment to adapting TUS to modern needs and technological advancements.

TUS datasets that have been used in the study:

- TUS, individual: occupant, household demographics
- TUS, daily: daily activity records of occupants, i.e. daily activity and presence schedules

3.2.1.1 Sample Design - Population

The HETUS aims to collect information about the resident population living in households within a country. The definition of "resident population" states individuals who have their usual residence in these private households. "Usual residence' is defined as the place where a person spends their daily time. This includes those who have lived at their place of residence for at least 12 months before the survey or those who moved there with the intention to stay for at least a year. In cases where this can't be established, legal or registered residence is used. For those living in multiple residences, the place where they spend most of the year is considered their usual residence. Special considerations are made for workers away from home, students, and children in shared custody. Thus, the framework is focusing on factors related to household contribution and benefits receipt to determine their usual residence.

The concept of a “household” extends to individuals sharing income or expenses, even if they are not related. In cases with ambiguous household classifications, the interviewee’s judgment is used. For addresses housing multiple households, the surveys aim to record data for all households, but the double counting is avoided for each occupant in the similar position. The concept of economic interest is applied to determine household residency. Thus, the process is based on the assignation of a household principal for the residential as a representative base for all its members.

3.2.1.2 Number of Diary Days

It uses two diary days category (or weekly day category), i.e. one weekday (Monday-Friday) and weekend-day (Saturday and Sunday). While employing a single diary day for an entire weekday can be sufficient. However, it is mostly not possible to get any idea of intra-weekday variations. HETUS categorizes the weeks based on the seasonal variations thus, these variations also increase the understanding of the occupant behavior. The general rule is that "the more diary days the better". Therefore, to achieve a robust representation of occupancy schedules, incorporating a broader range of weekly or daily variations is necessary.

3.2.1.3 Survey Forms (i.e. Questionnaire)

In the HETUS framework, three primary survey forms (i.e., questionnaire) are used which are the household, the individual, and the time-use diary. These surveys are targeted at household members aged 10 years and older. The household questionnaire is designed for face-to-face interviews with a knowledgeable household member, i.e. Pen-and-Paper Personal Interviews (PAPI). The individual questionnaire can be organized in person, by telephone, or by proxy if certain members are absent during the initial visit. This proxy application is arranged with the household principal.

The individual and household questionnaires are basic templates that can be adapted to country-specific norms and procedures, those used in labor force surveys. Especially, this is possible in the HETUS countries. While they include technical terms for clarity, these can be simplified, and the order of questions can be changed to suit practical needs. In addition, countries have the flexibility to introduce additional questions as needed.

Time	***What were you doing?*** *Record your main activity for each 10-minute period from 07.00 to 10.00!* Only one main activity on each line! Distinguish between travel and the activity that is the reason for travelling.	***What else were you doing?*** *Record the most important parallel activity.*	***Did you use a computer, smart device, internet, online tool, or similar technology or device for doing this?*** Yes	***Where were you?*** *Record the location or the mode of transport.* e.g. at home, at friends' home, at school, at workplace, in restaurant, in shop, on foot, on bicycle, in car, on motorbike, on bus, ...	***Were you alone or together with somebody you know?*** ***Mark "yes" by crossing***					
					Alone (or with unknown persons)	With other household members: Partner	Parent	Children (up to 17 years)	Other house-hold member	Other persons that you know
10:00-10:10	*Work*	*Coffe break*	☐	*Workplace*	☐	☐	☐	☐	☐	☒
10:10-10:20			☒		☒	☐	☐	☐	☐	☐
10:20-10:30			☒		☒	☐	☐	☐	☐	☐
10:30-10:40			☒		☒	☐	☐	☐	☐	☐
10:40-10:50		*Meeting with colleague*	☐		☐	☐	☐	☐	☐	☒
10:50-11:00		–"–	☐		☐	☐	☐	☐	☐	☒
11:00-11:10		–"–	☐		☐	☐	☐	☐	☐	☒
11:10-11:20			☒		☒	☐	☐	☐	☐	☐
11:20-11:30			☒		☒	☐	☐	☐	☐	☐
11:30-11:40	*Lunch break: had lunch*	*Talked with colleagues*	☐	*Canteen*	☐	☐	☐	☐	☐	☒
11:40-11:50	–"–	–"–	☐	–"–	☐	☐	☐	☐	☐	☒
11:50-12:00	–"–	*Checked the news*	☒	–"–	☐	☐	☐	☐	☐	☒
12:00-12:10	*Lunch break: went to supermar*	*Listened to music*	☒	*On foot*	☒	☐	☐	☐	☐	☐
12:10-12:20	*Lunch break: bought food*		☐	*Supermarket*	☒	☐	☐	☐	☐	☐
12:20-12:30	*Lunch break: went back to work*	*Listened to music*	☒	*On foot*	☒	☐	☐	☐	☐	☐
12:30-12:40	*Work*		☒	*Workplace*	☒	☐	☐	☐	☐	☐
12:40-12:50			☒		☒	☐	☐	☐	☐	☐
12:50-13:00			☒		☒	☐	☐	☐	☐	☐

Figure 3.3. Standardized Question format of HETUS

Figure 3.3 displays the structured questionnaire format used in the Harmonized European Time Use Survey (HETUS). The questionnaire format is detailed and provides detailed information on the participant's (i.e. occupant) activities, i.e. column of "occupant_activity". The survey records a participant's activities in ten-minute intervals. This recording includes primary and secondary activities, as well as the use of technological devices. However, in the current study, these features are ignored.

The location (i.e. column of “location”) of each activity is also captured with categorical options and space for specifics, e.g. as “Workplace” or “Canteen”. However, in this study, only residential presence was important thus all the exterior locations are grouped under the outside category. Thus, the location information’s are converted to binary information i.e. 1: inside, 0: outside. For instance, this observation of the daily activities reveals a diverse range of activities from work-related tasks to leisure. This can consist of lunch break activity, alongside transitions in location. It also provides side-information as the integration of technology or painting a detailed picture of an occupant’s time use as shown in Figure 3.3. This recording does not ignore the social interactions within a defined timeframe, mostly with family members if they are in the residential unit.

3.2.1.4 Activity Coding List

The Activity Coding List (ACL) in the HETUS is a comprehensive categorization system used to classify and code various activities reported by occupants in time use surveys (Eurostat, 2018). This list is structured to capture the full spectrum of daily activities, ranging from employment, to leisure, social life, and other miscellaneous tasks. Each activity is assigned a unique code, often at multiple levels (e.g. 1-digit, 2-digit, 3-digit codes). Thus, it provides both broad categorizations and detailed sub-categories. This coding system is crucial for the standardization and comparability of time use data across different countries and populations.

The ACL serves not just as a tool for systematically documenting how occupants spend their time but also enables the analysis and interpretation of time use patterns in relation to various socio-economic factors. The structured nature of the ACL allows for nuanced insights into daily life, influencing policy decisions and contributing to a deeper understanding of societal dynamics. However, in this study, ACL is used only to understand inside-home activities of the occupants in the residential buildings. Detailed socio-demographic aspects included in the ACL are not considered. The technicalities of this approach are elaborated on in the following subchapters.

3.2.2 The Selected Datasets: TUS and CENSUS

The TUS dataset is compiled by ISTAT. The main aim of the gather the daily activities and time use of occupants (e.g. work, leisure, caregiving, and domestic tasks) within Italy's residential building stock. By capturing data on time spent on various activities, the TUS helps in identifying social trends, gender disparities in time use, and the evolving needs of different population segments, especially at household scale. This dataset is particularly valuable for researchers and policymakers aiming to understand the implications of daily activities on societal well-being and economic productivity.

The CENSUS dataset collected by ISTAT represents a critical data for understanding various demographic and socio-economic aspects. The primary purpose of this dataset is to provide a comprehensive snapshot of the population at a given time (or specified years). This captures the information on population size and distribution, including age, sex, household composition, and housing characteristics. It also offers insights into education levels, employment status, and migration patterns. This extensive data collection serves as an invaluable resource for policymakers, researchers, and governmental agencies and it is conducted periodically.

Consequently, the following chapters highlight the process of refining and enhancing the datasets to ensure that they are more representative and effective for UBEM, focusing on the transformation from static population statistics to a dynamic, detailed occupancy dataset that accurately reflects energy consumption patterns in urban residential environments.

3.2.3 Time-Use-Survey Dataset (TUS)

The chapter describes an Explanatory Data Analysis of the TUS, which currently consist of TUS Daily (with only sequential data), and TUS Individual (ISTAT, 2018, 2022). In this process only manual refinement of the dataset and removal of non-essential columns are applied for targeted analysis.

The presented dataset is a comprehensive collection of time-series data, primarily focused on household dynamics and occupant activities. It includes a wide range of variables, providing a multi-dimensional view of the occupants' daily lives. The dataset contains the identifications for each household and its occupants, named as "Household_ID" and "Occupant_ID_in_HH". This level of detail enables in-depth analysis of occupant behaviors within the household context.

3.2.4 The CENSUS Dataset

The dataset is derived from the 15th Italian General Census of Population and Housing conducted in 2011 (ISTAT, 2013a). It is a representative sample of 1% of the national population. The census was carried out on March 31, 2014, in compliance with Framework Regulation 763 of 2008. Data was acquired through a bifurcated approach. A Computer-Assisted Personal Interviewing (CAPI) system facilitated online questionnaire responses, and a traditional Pen-and-Paper Personal Interview (PAPI) technique was employed for the paper-based questionnaire. The online method accounted for approximately 33.4% of the total responses, while the paper-based method accounted for the remaining 66.6%.

The population sampling strategy followed specific criteria that is targeting within municipalities with a minimum population of 20,000 or those classified as provincial capitals. A randomized and non-repeating selection of families was drawn from the personal data registry as of January 1, 2011. It is equivalent to 1% of the total Italian population.

The Census surveyed a variety of residential units, including families and single occupants living in surveyed dwellings, cohabiting individuals. The survey also accounted for individuals who were present temporarily on the census data.

Census datasets consist of the residential-related features. The housing data units consists of accommodations occupied by at least one permanent resident. During the census interval, the residential units were used as the foundational elements for data collection on residential characteristics, besides household compositions, and the demographics of resident individuals in Italy.

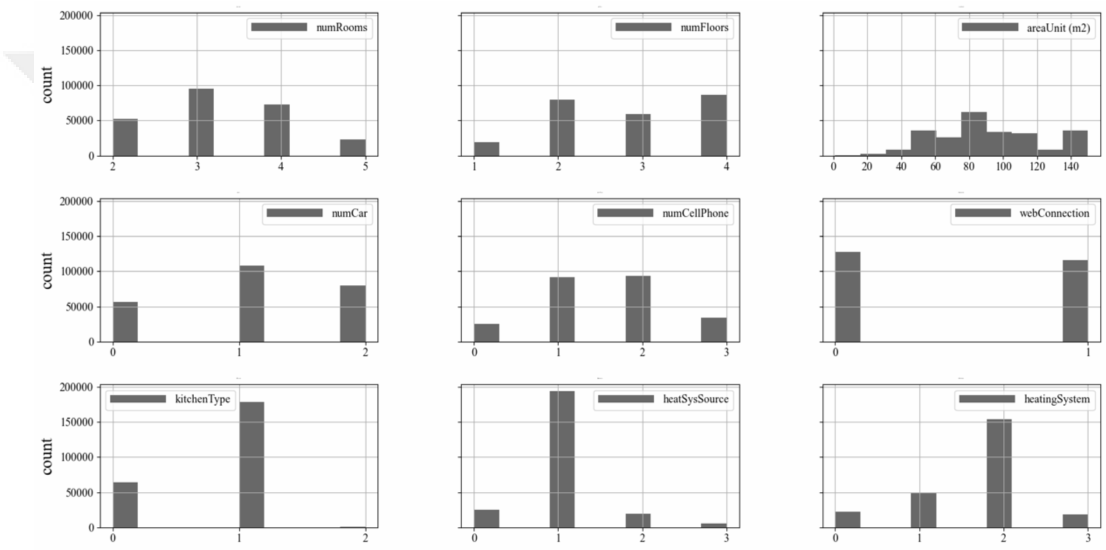

Figure 3.4. Distributions of UBEM-related columns in the CENSUS

Figure 3.4 displays histograms that show the distribution of residential attributes after removing outliers using a 25-75% interquartile range (IQR) process. The histograms indicate the number of data points in each category after the removal of outliers.

- “numRooms” indicates the prevalence of households with 2 to 5 rooms but in the original data, residential buildings with 1 room or 6 rooms are also preserved.
- “numFloors” shows the number of floors in residential buildings,

- “areaUnit (m^2)” provides insight into the of dwelling sizes.
- “numCar” shows the number of cars per household
- “numCellPhone” shows the number of cell phones per household
- The infrastructure attributes of the dwelling as “kitchenType”, indicate the frequency of homes with different kitchen configurations.
- “heatSysSource” and “heatingSystem” show the types of heating systems and their sources of energy.
- “webConnection” indicates the presence or absence of internet connectivity in these homes. This data was more important at the time of the data collection of CENSUSES 2011 due to average web usage adaptation in the residential buildings.

These histograms provide a quantitative overview of the residential characteristics of the dataset. Understanding these characteristics is critical for analyzing patterns in energy demand and occupant behavior, especially after removing statistical outliers.

The dataset comprises both census population and census housing components. For this study, a merged subset of these census datasets has been prepared. Then only the number of rooms and household size columns are extracted for use in the developed occupancy model. However, this combined census dataset will be employed more extensively to encompass broader areas.

3.2.5 Data Analysis of TUS temporal

The *TUS temp,* i.e. TUS temporal, dataset is the combination of TUS individual and *TUS daily* datasets derived from the ISTAT. Thus, the *TUS augmented* consisting of the properties of occupants, households, residentials as non-sequential columns and hourly presence, activity schedules as temporal dataset. Non-temporal but synchronized with temporal columns as similarly executed in B. Yang et al. (2021).

Structure:

Each category belongs to the occupants in the residential household and thus, categories are related to each other. That means that each season one-week schedule is available, and each weekly schedule consists of three different daily schedules (1: weekdays, 2: Saturday, 3: Sunday). These will be denoted as weekday categories (or it is evaluated as "dairy days" in HETUS format) to represent a week. Each non-sequential column, e.g. "region", is synchronized with sequential columns for the integrity, e.g. "occupant_activity".

However, as a limitation of the TUS datasets is that each occupant is represented with only a single daily activity schedule. For example, *occupant x* might reflect a typical winter Sunday routine, whereas *occupant y* might correspond to a summer Saturday schedule. This is the combinarion of "months_season" and "week_or_weekend" columns.

Target Variables:

The dataset includes critical temporal columns such as "Occupant_Activity" and "location". "Occupant_Activity" is a categorical and integer-based column that crucially delineates the range of daily activities of the occupants. This column is of paramount importance as it categorically outlines the occupants' daily activities. "Location", on the other hand, indicates the activity's location, coded as binary and integer.

The Unique Structure of TUS Temporal

TUS temp dataset is a deviation from the usual sequential datasets that are frequently used for occupancy modeling. Unlike conventional datasets that primarily consist of temporal sequences, *TUS augmented* dataset integrates a hybrid model of both temporal and non-temporal characteristics. The non-temporal features consist of occupant demographics that present a static profile of household characteristics, including the number of occupants and their socio-economic attributes. This provides the foundation for the temporal dataset that captures the dynamism of daily activities across 30-minute intervals.

Each daily activity schedule in the *TUS augmented* dataset is organized into 48 rows that is capturing 24-hour day divided into half-hour intervals. This level of detail enables thorough mapping of occupant behavior without any overlap or discontinuity. This discrete daily encapsulation enables a precise and detailed analysis of occupant patterns and their subsequent energy demand profiles.

3.3 Occupancy Dataset for UBEM

The formation of an occupant-centric dataset is an important step in refining the UBEM for residential building stock. This comprehensive dataset unites a variety of critical elements to ensure a holistic and accurate representation of the urban building energy dynamics. Key components of this dataset include detailed specific weather patterns, and physical properties which is important for this compilation mentioned in section 3.2.4, e.g. the number of floors and residential units.

Occupancy-related schedules include gathering data on the number of people, household presence, activity schedules, and the use of equipment such as plug loads and lighting. Thermostat set points and corresponding schedules are also integrated to present real-life energy usage patterns. For this study, certain parameters as Domestic Hot Water (DHW) usage and occupant actions related to window schedules (i.e. occupant behavior related schedules) for ventilation are intentionally omitted from this dataset. Construction properties of buildings, e.g. U-values, Solar Heat Gain Coefficients (SHGC), and HVAC systems are included to provide an understanding of the building's thermal performance.

Together, these separate yet interrelated data sources constitute the foundation of the occupant focused UBEM dataset, providing essential detailed insights that are key to precise energy modeling and simulation.

3.3.1 Occupancy Modeling Related Components

Key datasets include the Italian Census Housing and Occupant Sub-Datasets (2011) and the Italian Time Use Survey (TUS) Individual and Daily Datasets (2013), all from ISTAT (ISTAT, 2013a, 2018). Together, these datasets provide a comprehensive daily representation of residential households in 19 different administrative regions of Italy, down to the sub-municipal level. The 15th Census of Population and Housing data warehouse, provides mixed details on Italy's demographic and social structure, as well as its residential building stock.

The TUS, an essential part of the "Multiscopo" (Ing. Multipurpose) family survey system, stands out for its detailed recording of occupants' time allocation at 10-minute intervals, providing insights into the organization of daily life, gender perspectives and the division of household roles. For this study, 10-minute intervals are updated with 30-minute intervals to facilitate the learning process of deep learning models. The survey methodology combines one-stage and two-stage sampling, ensuring representative estimates and covering a wide temporal range. This comprehensive approach allows for understanding of occupancy patterns and their impact on energy demand, which is crucial for UBEM in the Italian context.

Italian Census Housing Sub-Dataset, 2011, ISTAT

The Census Housing from ISTAT provides a comprehensive view of residential characteristics within Italy (ISTAT, 2013b). Key features are grouped into categories as accommodation ID, residential properties, e.g. including kitchen, floor area, room types, conditioning, and energy systems e.g. heating, and cooling systems, renewable energy sources, and vehicle properties.

In addition, communication features like mobile, phone, and internet availability are also included. The dataset's content covers the physical aspects of households and their energy setups, critical for urban building energy modeling. Pre-processing steps include renaming column headers, cleaning and merging columns with redundant or singular information, manual imputation based on column interrelations, and assessing feature importance. This rich dataset provides a solid foundation for understanding the spatial and infrastructural dynamics of Italian residences, pivotal for informed energy modeling and policy development.

Italian Census Occupant Sub-Dataset, 2011, ISTAT

The “Italian Census Occupant (Household) Sub-Dataset, 2011 from ISTAT offers critical insights into household demographics in Italy. This dataset, encompassing roughly 276,000 rows and 51 columns, includes data on household locations, personal characteristics of occupants, and their work and school attendance frequencies.

Italian TUS Individual Sub-Dataset, 2013, ISTAT

The "Italian TUS SUBSET dataset, 2013" from ISTAT is a comprehensive and intricate collection of household and occupant-related data, encompassing around 45,000 rows and 267 columns. This subset of the Time Use Survey is conducted every five years, targeting a random sample of approximately 34,000 households across 650 diverse Italian municipalities. It is a part of ISTAT’s larger "Multipurpose” survey system on families and is particularly notable for its *dairy-based* methodology. Respondents document their daily activities, commutes, and interactions in 10-minute intervals over a 24-hour period. This detailed recording offers invaluable insights into the daily lives and routines of Italian residents, providing a nuanced understanding of societal and familial role divisions, especially from a gender perspective.

The dataset consists of many features covering socio-demographic characteristics, family economic situations, health conditions, daily activities, and perceptions of life. It includes detailed information on household properties e.g. family typology and location, and occupant properties e.g. gender, age, marital status, education, and employment details. Dataset captures data on more subjective aspects like satisfaction with daily life and opinions on gender roles, although these were not a focus in this study. The TUS subset acts as a crucial link between census data and other TUS datasets, detailing temporal patterns of home departure and return for work or education. This rich dataset thus serves as an essential tool for analyzing the intricate dynamics of Italian households, contributing significantly to our understanding of urban residential patterns and their implications for occupancy modeling.

The non-sequential TUS individual dataset consists of a range of categorical and integer-based columns, each capturing distinct demographic and socio-economic attributes occupants. “Region” presents the location of a household within Italy, categorized uniquely into five regions, e.g. north, south, etc. “Number Family Members” quantifies the household size, ranging from one to six members. “Family Typology” classifies households into four distinct types as reflecting the family structure, e.g. "Lonely person", "Set of relatives / Relatives and others / Non-relatives", "Childless couples", "Couples with children", "Mother with children", "Father with children", "Families with two or more nuclei", "Parent with non-single/single children". “Age Classes” categorizes occupants into six age groups, providing insights into the age distribution within the population. “Education Degree” captures the educational attainment of occupants, i.e. "No title", "Elementary school certificate", "Middle school certificate", "Diploma", "Graduate or postgraduate". “Employment Status” details the employment scenario of occupants, divided into six categories, i.e. "Employed", "Looking for first job", "Unemployed", "Retired", "Student", "Housewife", "In another condition". These columns collectively facilitate a nuanced understanding of the Italian household dynamics. In addition to these features, “Household_ID” and “Occupant_ID_in_HH” also exist in the dataset similarly with TUS Daily.

Table 3.1 A sample from TUS Individual dataset

Household_ID	*Occupant_ID_in_HH*	*Region*	*Number Family Members*	*Family Typology*	*Age Classes*	*Employment status*	*Gender*	*Education Degree*
3	1	3	1	2	3	1	2	2
4	1	1	1	2	3	1	1	3
5	1	1	2	4	2	1	2	2
5	2	1	2	4	1	6	2	2
6	1	4	2	5	4	1	1	1
6	2	4	2	5	4	1	2	2
7	1	3	2	5	5	1	1	3
7	2	3	2	5	5	3	2	2
9	1	1	2	4	5	3	2	1
9	2	1	2	4	3	1	2	2
10	1	1	3	4	6	3	2	1
10	2	1	3	4	5	2	2	2
10	3	1	3	4	4	1	2	2
13	1	4	4	3	6	6	2	1
13	3	4	4	3	1	6	2	1
13	4	4	4	3	3	5	2	2

A statistical measures offer an overview of the data's distribution and variability, which is important for understanding general patterns and outliers in the dataset. Additionally, unique identification is included. The field "unique_id" serves as a unique identifier for each data entry, ensuring the accuracy and integrity of the dataset. It is the combination of household and occupant identification; it is used in the transitory processes thus it is not shown in the tables.

Table 3.2 A statistical description of TUS Daily with only sequential features

	months_season	*week_or_weekend*	*hourStart_Activity*	*hourEnd_Activity*	*Occupant_Activity*	*location*
mean	2.43	1.966	13.760	13.766	388.438	0.657
std	1.123	0.809	5.617	5.614	346.718	0.474
min	1.0	1.0	0.0	0.0	11.0	0.0
%25	1.0	1.0	9.0	9.0	31.0	0.0
%50	2.0	2.0	14.0	14.0	321.0	1.0
%75	3.0	3.0	19.0	19.0	736.0	1.0
max	4.0	3.0	23.0	23.0	997.0	1.0

Italian TUS Daily Sub-Dataset, 2013, ISTAT

"Time Use Survey" is a survey that covers the resident population in private households. The survey is executed by interviewing a sample of 20.000 households and 50.000 people (P.A.P.I. technique, Pen-and-Paper Personal Interviews). *TUS Daily* dataset compose of the hourly resolution daily occupancy and activity schedules for every member of the household, approximately 50000 people. The daily actions are recorded for about 24-hours however, the total number of actions is not the same for every occupant, it varies between 6 to 82 actions per day.

Table 3.3 A sample from TUS Daily dataset

Household_ID	*Occupant_ID_in_HH*	*months_season*	*week_or_weekend*	*hourStart_Activity*	*hourEnd_Activity*	*Occupant_Activity*	*location*
427	1	3	2	0	0.5	111	0
427	1	3	2	0.5	1	111	0
427	1	3	2	1	1.5	111	0
427	1	3	2	1.5	2	111	0
427	1	3	2	2	2.5	111	0
427	1	3	2	2.5	3	111	0
427	1	3	2	3	3.5	111	0
427	1	3	2	3.5	4	111	0
427	1	3	2	4	4.5	111	0
427	1	3	2	4.5	5	111	0
427	1	3	2	5	5.5	111	0
427	1	3	2	6	6.5	111	0
427	1	3	2	6.5	7	111	0
427	1	3	2	7	7.5	111	0
427	1	3	2	7.5	8	111	0
427	1	3	2	8	8.5	111	0
427	1	3	2	8.5	9	111	0
427	1	3	2	9	9.5	821	1
427	1	3	2	9.5	10	821	1
427	1	3	2	10	10.5	11	1

The total row number is more than 1MM. for the deep learning is implemented in this study, the daily action number is limited to 48-number as representing each 30-

minute interval of the day, the variables “months_season”, “week_or_weekend”, “hourStart_Activity”, and “hourEnd_Activity” introduce spatial and temporal dimensions. These features have different values over time and in relation to the physical space of the home. The variables “Occupant_Activity” and “location” provide insights into the specific activities undertaken by the occupants and their location within the household. This data is directly representing daily routines and mobility patterns within the residential buildings.

3.4 Occupancy Modeling

For Occupancy Modeling for UBEM in this thesis, a comprehensive approach is adopted which is encompassing the collection of a diverse range of data, e.g. from occupant demographics to activities. Data is pre-processed to ensure suitable format and quality for the subsequent modeling steps. Detailed profiles of typical occupants in various residential buildings are developed for reflecting diverse behaviors and lifestyles. These profiles are utilized to simulate occupancy patterns, forming the basis occupancy information for UBEM.

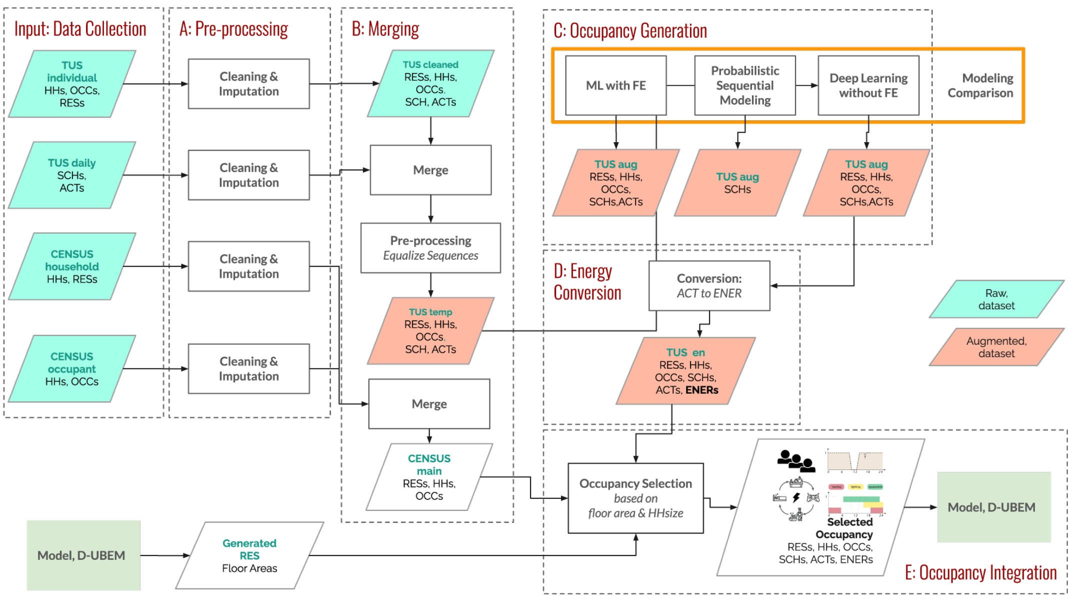

Figure 3.5. Occupancy Modeling Over Datasets & Steps

Figure 3.5 illustrates the process from data collection to the final generation of occupancy information. The process is multidimensional, integrating data cleaning, augmentation (for imputation), and advanced machine learning techniques to generate occupancy information including occupant and household profiles, presence schedules, daily activity schedules, energy use schedules. In summary, this methodology captures the complexity of occupant behavior in the built environment and translates it into actionable data that serves as a cornerstone for energy modeling.

Input Datasets: The proposed methodology utilizes with four primary datasets: two from the Time Use Surveys (TUS) (ISTAT, 2022), which provide individual and daily schedules (HHs, OCCs, SCHs, ACTs). In parallel, two datasets come from the Census (ISTAT, 2013a), which provide the connection between residential floor area and household size. TUS is the main dataset of the process and CENSUS is used while assigning the households to residential units.

Pre-processing: Each dataset undergoes preprocessing, which ensures data quality and fills in gaps to maintain dataset integrity, i.e. cleaning, manual imputation based on the missing columns.

Merging: After initial cleaning steps, the datasets go through merging phase using the combination of TUS individual and TUS daily to obtain TUS temporal. This is basically the merging the sequential and non-sequential dataset formation. In parallel, CENSUS main is formed by the merging of CENSUS housing and CENSUS occupant.

Occupancy Modeling: <u>The occupancy generation phase is multifaceted by employing three different approaches i.e., conventional method (CM) as ML without feature engineering, stochastic method (SM) using Markov models and end-to-end deep learning method (EM) using deep learning.</u> Approaches are applied for the imputation of the sequential data for each occupant e.g., occupant activity and activity location.

Energy-usage related Schedules: It is a critical step where activity data (ACT) is transformed into energy usage metrics (ENERs) i.e. lighting energy use and equipment energy use schedules based on presence schedule using information of activity location. <u>This dataset provides a detailed temporal view of when occupants interact with their environment and how these interactions translate into energy demand.</u>

Occupancy Integration: The culmination of this process is the generation of a dataset of Occupancy information i.e., Households (HHs), Occupants (OCCs), Activities (ACTs), Schedules (SCHs), and Energy Use Schedules (ENERs).

3.4.1 Limitations of Occupancy datasets

The limitations of the conversion of datasets for deep learning processes necessitate specific strategies to prepare the datasets effectively. These steps are vital in streamlining the datasets, thereby facilitating a more robust and accurate deep learning process. The challenges identified are as follows:

- <u>Challenge of Limited Daily Data in TUS:</u> The TUS dataset presents a limitation with only a single daily schedule available per occupant. To overcome this, schedule augmentation (i.e. expansion) techniques are employed to extrapolate more detailed daily patterns.
- <u>Occupancy-Related Discrepancies:</u> TUS dataset shows inconsistencies as a mismatch between columns i.e., the number of family members and the number of occupants. Thus, adjustments are necessary to ensure data accuracy
- <u>Integration Issues in Dataset Merging:</u> Integration issues arising from the merging of sub-datasets, particularly when dwelling identifiers are missing, the methodology employs corrective measures.
- <u>Selective Column Filtering:</u> For columns that are either irrelevant or contain significant missing data, e.g. as the number of cars or the web connection of

the residence. The essence of selective filtering is to strategically exclude extraneous details while focusing on critical elements, e.g. population distribution, living arrangements, and housing structures relevant to urban residential energy modeling.

- <u>Alignment of Demographic Categories:</u> Ensuring consistency and accuracy in the harmonized dataset involves aligning demographic categories to eliminate discrepancies, e.g. age groups and occupation types.

3.4.1.1 The Purpose of Augmentation for TUS Dataset

The raw dataset provides a single daily activity schedule for each occupant which is categorized by season (ie. Winter, Spring, Summer, Autumn) and weekday (ie. Weekdays, Saturday, Sunday) columns. The goal is to expand the dataset to include 12 different schedules per occupant, covering each season and weekday type to provide a more comprehensive view of daily routine of occupants. This expanded data set is consisting of both non-temporal (demographic) and temporal (daily activities in 30-minute intervals) data elements.

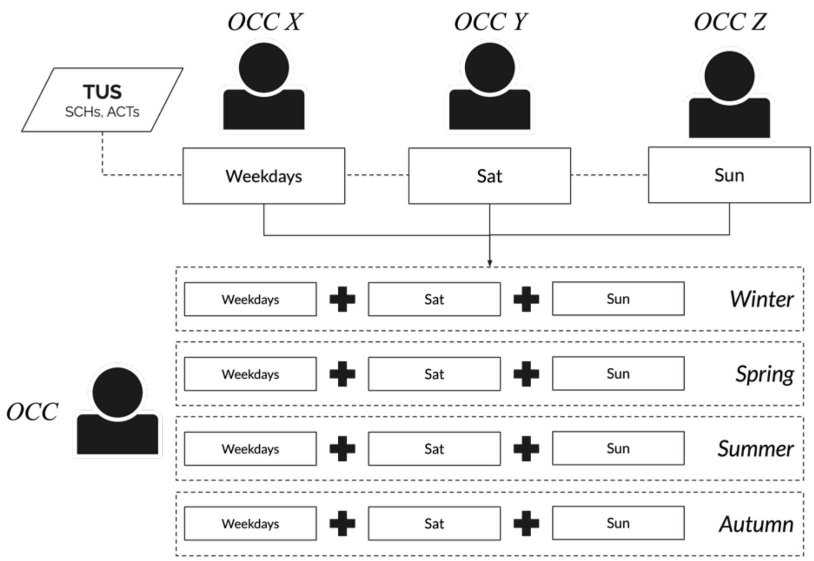

Figure 3.6. The rationale behind the augmentation of the occupancy dataset

3.4.2 Augmentation of Sequential Occupancy Information

Three methodologies are examined comparatively for augmenting the TUS temporal dataset (Figure 3.7)

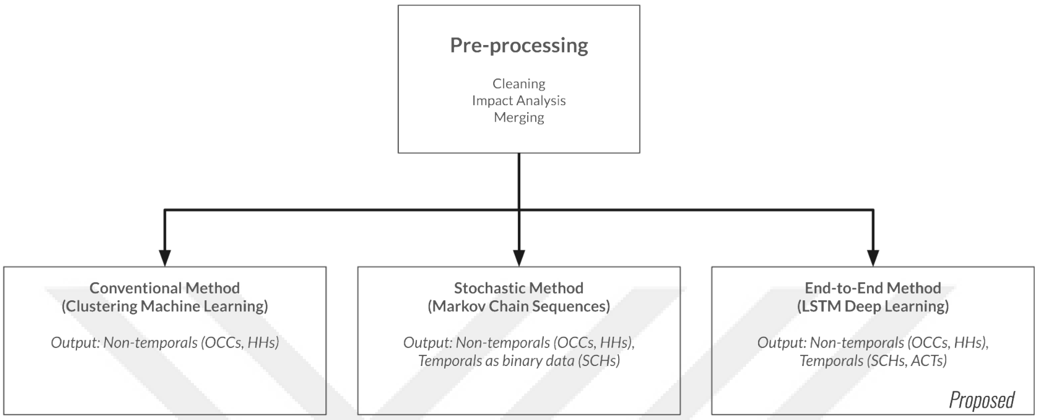

Figure 3.7. Methodologies of Occupancy Modeling

The first is a Conventional Method (CM) using machine learning with feature engineering which is focusing on occupant demographics and employing clustering algorithms to identify typical patterns. This archetype-based clustering predicts and completes activity schedules based on known demographic characteristics. The second method is a Stochastic Method (SM) utilizing Markov-Chain processes. Stochastic Methods are preferred from many researchers for mostly sequential presence schedule generation. SM generates sequential presence schedules for incorporating demographic information to enhance modeling capacity. The third method is an End-to-End Deep Learning Method (EDM) that processes and classifies both temporal and non-temporal features without explicit FE. EM is creating a comprehensive depiction of daily activities and locations for occupants across various seasons and days. This model enhances the dataset by representing realistic yearly schedules through the classification of 48 distinct time intervals per day.

LSTM approach offers automated management of large datasets with advanced predictive capabilities. The clustering technique is noted for its transparency and ease of application, despite its limitations in capturing the full complexity of daily activities. This comparative analysis highlights the trade-offs between interpretability and predictive power in urban residential energy modeling.

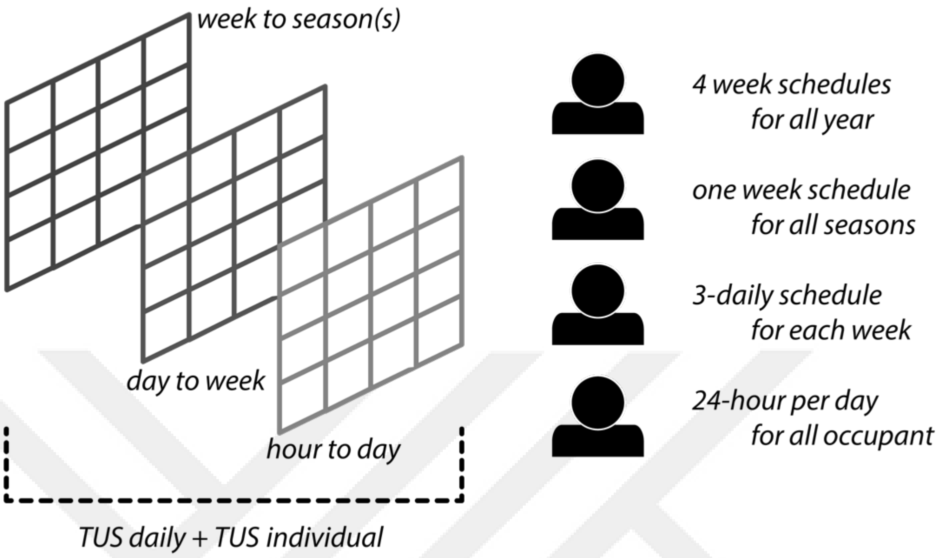

Figure 3.8. Temporal hierarchy of occupancy dataset

Figure 3.8 illustrates a hierarchical temporal structure of augmented occupancy datasets. It consists of three segments that correspond to different temporal scales: macro (week to season), mid (day to week), and micro (hour today). It is possible to distinguish between seasonal, weekly, and daily variations in behavior by segmenting time in this way. These variations impact residential energy usage.

3.4.2.1 Conventional Method (CM): Machine Learning with FE

The primary objective is to use archetype profiling to analyze demographic data, as shown in Figure 3.9. This involves identifying typical patterns or archetypes within the demographics, and then grouping similar archetypes together. The next step is data augmentation. It aims to fill missing data about daily activities of residential occupants. This filling process is based on the seasonal and weekday (i.e. diary days) parameters, i.e. "Months_season", "Week_or_weekend".

Archetype characterization is first step that is discerning patterns within occupant demographic, i.e. “Region”, “Number of Family Members”, “Family Typology”, “Age Classes”, “Employment Status”, “Gender”, and “Education Degree”. This process results in the formation of distinct archetypes. Occupants possessing incomplete daily activity schedules are identified and assigned to a cluster corresponding to their demographic archetype. For instance, an occupant (referred to as “Occupant Z”) with recorded activities for a summer Sunday but lacking data for other days. Thus, clustering is applied and several clusters are emerged. Occupant Z is under the cluster 5, based on their archetype. Daily activity data from other occupants within the same cluster is then utilized to complete the missing schedules of “Occupant Z”. This cluster-based approach ensures the consistent imputation of missing information.

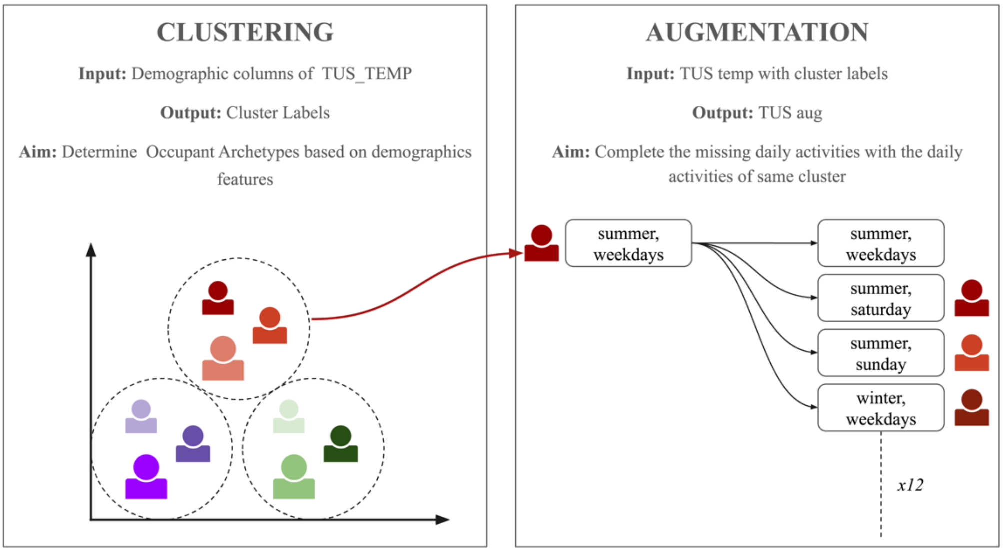

Figure 3.9. The main process of ML with Feature Engineering

The raw TUS dataset consist of a single daily activity schedule for each occupant, with seasonal and weekly variations being random. In contrast, the augmented TUS dataset provides each occupant with twelve distinct daily activity schedules. This also encompass three days per week across all four seasons.

Selecting the Optimal Clustering Approach

The process involves preprocessing features for better representation in the clustering algorithm and applying a hierarchical agglomerative clustering algorithm. This method is chosen for its insightful representation of occupant demographics. K-Means clustering is a centroid-based method. K-Means is generally used for numerical data and assumes spherical clusters which may not be suitable for mixed data types, e.g., TUS. DBSCAN is a density-based method. While effective in identifying outliers and not requiring pre-specification of clusters, the technique can be sensitive to parameter choices and may struggle with clusters of varying densities. Hierarchical Clustering is an unsupervised machine learning method that organizes data into a hierarchical structure. This is more adaptable to different data types and uses various distance metrics, making it suitable for mixed variables in our dataset. Thus, Hierarchical Agglomerative Clustering is recommended as the initial approach due to its compatibility with mixed data types and flexibility in handling various distance metrics.

Evaluating the Performance of the Model

Evaluating the performance of the Clustering involves method requires nuanced considerations. For supervised learning models, there are well-defined accuracy measures exist, however, for clustering (i.e. unsupervised learning) evaluations are more complex due to the absence of known ground truth labels, existing data. Clusters can be evaluated using different metrics. Using multiple metrics makes the evaluation better. The Silhouette Score measures the similarity of an individual data point to its own cluster compared to neighboring clusters. Its values are ranging from -1 to +1. Higher scores indicate better fit to the assigned cluster and less similarity with neighboring clusters. This provides insights into cluster cohesion and separation. The Calinski-Harabasz Index (i.e. Variance Ratio Criterion), scores clustering quality based on the balance between intra-cluster similarity and inter-cluster dissimilarity. High scores in this index suggest distinct clusters. Lastly, The

Davies-Bouldin Index calculates the average similarity between each cluster and its nearest counterpart. The lower values are indicating more distinct clustering. It is characterized by reduced overlap and clear distinction between clusters. This index assists in evaluating the compactness and separation of clusters by considering their relative similarity.

Hyper-parameter Tuning with Grid Search Methodology

In agglomerative clustering, tuning hyper-parameters like the linkage function and distance metric is crucial for accurate clustering. A grid search methodology is used to explore a range of parameters for enhanced performance. The process clusters demographic archetypes based on occupant characteristics and imputes missing daily activity data using “months_season” and “week_or_weekend” columns. Grid search iterates over values for the number of clusters ([10, 20, 30, 40, 50]) with the parameters of linkage methods (["ward", "single", "complete", "average"]), and distance metrics (["euclidean", "manhattan", "mahalanobis"]). Each iteration applies the Agglomerative Clustering model with different parameter combinations, resulting in various clustering solutions and ensuring the selection of the most suitable configuration for the dataset.

Application in the Context of TUS Temporal Dataset

In the application of agglomerative clustering to the TUS temp dataset, a two-stage tuning process is employed to optimize the representation of demographic archetypes. Initially, in the first phase, the method is applied to the dataset with a wide range of clusters (10 to 50), this is a broader exploration for the clustering space. The aim here is to identify a reasonable range for the number of clusters that accurately represents the demographic archetypes, evaluating performances using metrics, i.e. the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. These metrics are explained in the previous chapters.

Secondly, the refinement phase comes for a detailed scanning of the potential parameters ranges that are defined by initial phase. In this phase, the range of clusters is narrowed down to between 20 and 30. This step is aiming to enhance the precision of the clustering and thereby improving the quality of the demographic archetype characterization. This systematic approach ensures a comprehensive and effective analysis of the data. In the result chapter, the two process is explained with quantitative results.

3.4.2.2 Stochastic Method (SM): Markov-Chain Sequences

The main purpose of this part is the application of the Markov-chain process with different parameters i.e. increasing the modeling capacity with different Markov order levels and inserting the demographics into the modeling process.

Markov Chains offer a foundational probabilistic approach for modeling sequential data for residential occupancy schedules. The probability of each event depends on the state attained in the previous event(s). This property is called as the *Markov Property* in which future states are independent of past states.. If to consider ***S*** as a finite set of states, the MC can be represented as a sequence of random variables $X_1, X_2, X_3, \ldots.$ with the Markov Property. The transition probability is the likelihood of transitioning from one state to another. It is defined as $P_{ij} = P(X_{n+1} = j \mid X_n = i)$. P_{ij} represents the probability of transitioning from state ***i*** to state ***j***. In the context of daily presence generation in residential buildings, each state in the Markov Chain could represent presence status, i.e. present or absent. Then, the transition probabilities would capture the likelihood of shifting from one activity to another over time.

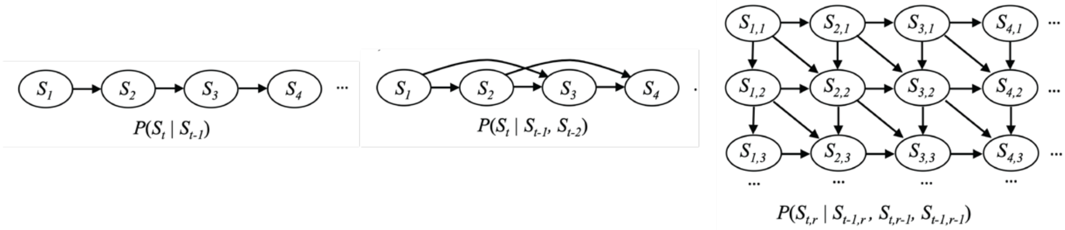

Figure 3.10. The Markov-chain state representations; (left) first order, (middle) second order, (right) higher order (Snodgrass & Ontañón, 2014)

Markov Chains can be categorized by the order of dependence that the model assumes between consecutive states.

- A first order (i.e. single order) Markov Chain assumes that the future state depends only on the current state, not on the sequence of events that preceded it, as shown in Figure 3.10 for the second and higher order dependence. This approach is computationally efficient. However, this may not always capture the complexity of processes where future states may depend on multiple preceding states.
- A second-order Markov Chain posits that the probability of transitioning to a future state is dependent upon the two previous states. This can be formalized as $P(X_{n+1} = j \mid X_n = i, X_{n-1} = h)$, where the transition probability accounts for the last two states, ***i*** and ***h***, before moving to state ***j*** (i.e. current state).
- Higher-order Markov Chains can evaluate broader context which can significantly enhance the predictive accuracy for complex sequences. Thus, Higher-order MC can be employed for more complex situations. On the other hand, there is a trade-off in terms of computational complexity. Because there is a need for larger datasets to accurately estimate the transition probabilities for higher order models.

Comparison Metrics for Markov-Chain Evaluation

This study uses three metrics to evaluate how well Markov-chain models predict presence schedules in the TUS dataset, each providing unique insights into the models' performance.

The Hamming Distance (HM) metric is selected to ascertain the dissimilarity between two binary sequences by counting the divergent positions, i.e., 30-min intervals same as LSTM. A lower value is indicative of increasing similarity between the two schedules as the ideal value being zero. The greater the HM means the more pronounced is the disparity between the generated and actual schedules. This metric is particularly beneficial in quantifying the exact count of intervals where the predicted schedule diverges from the observed reality.

The Jaccard Similarity (JM) metric is useful in assessing the similarity and diversity of sample sets. It compares how often two schedules agree (both having a 0 or a 1 at the same time) to the total number of time slots considered. The metric scales from 0 (no similarity) to 1 (perfect similarity). Jaccard Similarity is particularly valuable for explaining the extent of overlap between the schedules by offering a normalized and proportionate perspective of similarity.

Accuracy (Acc) serves as a standard metric as traditionally used in evaluating classification models. In parallel, it is applicable for comparing binary sequences. It measures the proportion of compatibility of predicted states in terms of the total number of states. Therefore, this indicates the extent of agreement between the predicted and actual schedules. The metric varies between 0 (no agreement) and 1 (complete agreement) with higher values signaling a more accurate prediction of the schedule.

In short, HM concentrates on quantifying the exact number of discrepancies, JM provides a proportionate assessment of the overlap, and Acc offers a direct measurement of the predictive concordance.

Limitations of MC for Occupancy Schedule Generation

Long-term MC models are stochastic models that assumes that state changes over time-based according to what happened in the previous activities. However, social practice theory suggests people's daily and weekly schedules are a complete picture (Torriti, 2017). This theory composes of these explanations that practices are done by groups of people and people behave repeatedly. These practices are connected and rely on each other. For instance, people wear clothes before leaving the home, not after. This means we need models that can show how these practices are linked and influence each other and understand the all-day patterns, i.e. 24 hours and 48-rows in the case of this study. However, Markov process only consider look at past activities to predict the future.

Models should understand how activities are part of our social life to really get how they change over time according to people's characteristics (Walker, 2014). Most people live their lives in daily patterns, e.g. work hours, mealtimes, and other routines. These routines help sync up social activities and patterns of demand.

Its "memoryless" nature: MC conceptualize each state as a segment of a stochastic process where the future state is only influenced by the present state. This presents the 'memoryless' nature inherent to the Markov property. This method is adequate for analyzing collective behavior however, it is inadequate for capturing extensive, individual behavioral patterns over longer periods. The inability to retain 'memory' beyond the immediate situation limits its performance in predicting complex life patterns spanning a longer period, for example, 30-min intervals of a single day.

Ignoring the household during activity classification: Although Markov models are useful in understanding broad population behaviors however, their structural constraints limit their applicability for individual-level behavior analysis. Because this type of modeling focus on long-term dependencies and varied behavior patterns.

Assumption-based FE: TUS dataset is the combination of both sequential and non-sequential data. Implementing these diverse data types in a Markov chain model necessitates extensive feature engineering. This is because Markov chains are designed primarily to predict or classify the states without looking interdependent relationships. An example FE is “activity” and “location “columns must be amalgamated into combined states like 'sleep at home', 'study in home', or 'study outside'. Similarly, the time of day needs to be integrated into these states. The intensive feature engineering process involves making assumptions that can significantly influence the accuracy of activity predictions. Since each assumption introduces potential bias or error. This limitation poses a need for more advanced, flexible models for handling complex datasets with both sequential and non-sequential data.

For “stateSequence=1” (35 activities):

- Single activity states: 35
- Transition matrix size: 35 × 35

For “stateSequence=2” (pairs of 35 activities):

- Pairs of activities: $35^2 = 1225$
- Transition matrix size: 1225 × 1225

Figure 3.11. The transition matrix computation of the activity schedules

Computation Cost of the MC: The computational implications of the MC model's order on the state transition probabilities could be limited. Given that occupancy schedules compose of long-term dependencies related to sequential format of daily activities and presence patterns, thus, detailed observation becomes imperative. Figure 3.11 illustrates an occupant engaging in 35 distinct activities, then, escalating the order of the Markov model exponentially increases the computational cost of the classification. Any enhancement in classification performance is likely to result in an exponential increase in computational demand as TUS dataset encompasses 146 distinct activities. Consequently, this puts a limit for the MC models in terms of generating activity schedules especially given the substantial computational resources required for such an endeavor.

3.4.2.3 Deep Learning Application for Augmentation Process

The selection of an advanced deep learning model as main model in this research is fundamentally driven by the need to capture the inherent sequential structure of activity data within the occupancy dataset. Traditional occupancy modeling methods often fail to capture the dynamic and diverse nature of occupants' daily routines due to focus on single-day activity logs or lack of dependent modeling of occupancy components. In the original dataset of TUS, there is only one daily activity schefule per occupant thus, to represent the all year long occupancy, the deep learning model can augment the one daily schedule for all seasons and weekday categories.

To overcome the limitations of the previous modeling approaches, this research innovatively integrates occupant-specific data, including non-sequential demographic and residential characteristics, with sequential daily activity logs. The model's ability to generate integrated multi-day activity schedules for an entire year that reflect true occupant diversity and behavior represents a novel approach to occupancy modeling.

3.4.2.4 End-to-End Deep Learning Method (EDM): Multi-task

This chapter describes the approach used to extend the TUS temp dataset using LSTM for the complete occupancy datasets for all year. LSTM is formed to process daily activity and presence schedules. The dataset is stratified into 48 discrete intervals, each corresponding to a 30-minute interval within a 24-hour period. This temporal resolution allows for a comprehensive representation of an occupant's daily routine in terms of activity, timing and the location information.

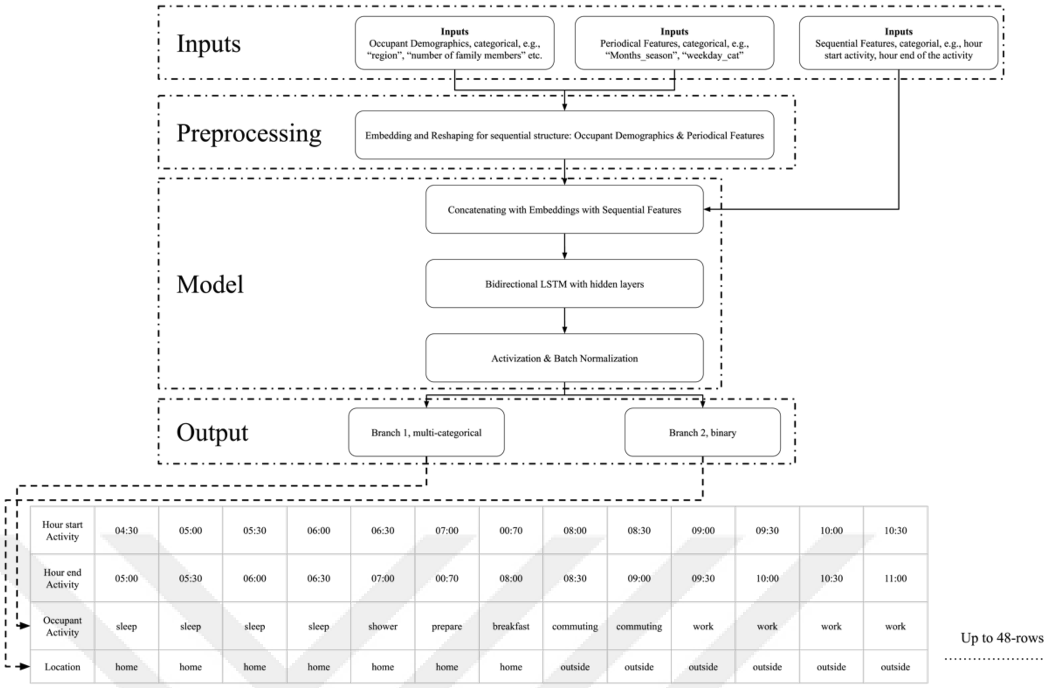

Figure 3.12. LSTM, model architecture

EM utilizes the TUS temp dataset for multi-task classification and the model classify each hour activity by capturing the intricate interplay between an individual's demographic profile and their daily activities. TUS temp dataset is an amalgamation of both non-sequential and sequeantial elements. Non-temporal data includes categorical and ordinal variables that denote demographic information, seasons, and days of the week. Temporal data is represented through ordinal variables that mark the beginning and end of each activity within the 30-minute intervals.

This research introduces a novel modification to the LSTM framework. Conventional LSTM applications often emphasize the single final output of sequential data, this study seeks to implement a multi-categorical classification for each occupant interval, that means, for each classification consist of 48-rows of activities as shown in Figure 3.12. The study assumes that each time segment holds unique behavioral patterns, which benefit from independent analysis. Accordingly, a multi-categorical classification is implemented, generating 48 distinct outputs for each interval. Additionally, an embedding structure is incorporated to integrate non-sequential information, such as demographic data, into the sequential input of the LSTM. It is believed that these combined modifications enhance the LSTM model's ability to identify and process the complex subtleties inherent in interval-based data.

LSTM (Long Short-Term Memory) Models

LSTM is a type of recurrent neural network (RNN) which is designed to capture and retain information across extended sequences of data. LSTMs were developed to address the limitations of traditional RNNs which is struggling to maintain stable gradients over long sequences, i.e., vanishing gradient problem.

LSTMs use *input*, *output*, and *forget gates* to control information flow which can retain important data and discarding non-critical information. This attribute enhances their predictive capabilities that is making LSTM as ideal method for time-series prediction, natural language processing, and sequence generation. Their ability to grasp context over time makes LSTMs essential in deep learning research and practice.

Multitask Learning in Occupancy Schedule Classification

This approach assumes that combining these variables creates a more effective single description for classifying things i.e., "occupant activity," "location". The methodology of multi-task classification as follows:

<u>Conjoint Representation Learning:</u> This shared representation is anticipated to encapsulate the commonalities within the dataset, thereby enhancing the model's generalization capabilities across the varied tasks.

<u>Inductive Bias as Regularization:</u> The adoption of multitask learning inherently introduces an inductive bias, which can bee seen as form of regularization. Doing multiple tasks at the same time helps the network avoid learning from random errors and irrelevant details in the training data, because the activity representation is now supported with the location information.

Inter-Task Knowledge Transfer: Multitask learning creates an environment where information can be shared between tasks. In the proposed model, it's implemented as the learning related attributes (like "location" and "occupant activity") at the same time can improve the ability to classify any one of them.

Computational Efficiency: Multitask learning affords an improvement in computational efficiency by yielding multiple output classifications within a singular forward pass of the network. This efficiency does not come at the expense of classification accuracy by leveraging the shared structure within the network. There are two reasons why adding location data doesn't significantly slow down the model. First, the model is designed to handle multiple tasks at once. Second, location data is simple (binary), unlike the more complex occupant activity data.

Data-Preprocessing of TUS Temporal for LSTM

The preprocessing function is structured into distinct segments:

Stratified Splitting Function:
- Combine 'months season' and *'week_or_weekend'* into *'strata'*.
- Perform stratified shuffle split on 'strata' to obtain train and test IDs.

Main Data Preprocessing Function:
- Initial Inspection: Filter data based on group size criteria.
- Integer Encoding: Convert categorical columns to numerical codes.
- Temporal Encoding: Transform *'hourStart_Activity'* and *'hourEnd_Activity'* using *sinusoidal* and *cosinusoidal* functions.
- Normalization: Apply scaling to selected columns.
- Data Splitting:
    - Stratified split to ensure balanced representation in train, validation, and test sets.
    - Merge data with corresponding IDs for each set.
- One-Hot Encoding: Convert 'Occupant Activity' to one-hot format.
- Reshaping for LSTM: Adjust data shape to fit LSTM requirements.
- Return processed datasets and one-hot encoder.

Figure 3.13. Pseudocode of Data Pre-processing of CENSUS, RAW Dataset

Stratified Data Splitting: The process is designed to create training and test subsets from the dataset while maintaining a uniform distribution of two categorical features i.e., "months_season" and "week_or_weekend". This stratification ensures that seasonal and weekly patterns are evenly represented across the subsets.

Initial Data Formatting and Cleaning: The data is grouped based on key identifiers which are "Household_ID" for household determination, "Occupant_ID_in_HH" for occupant order in the household, "months_season" for seasonal variation and, "week_or_weekend" for weekday categories. After formatting, a filtering step is executed for the schedules that do not match with 48-rows log setup. Thus, data integrity and consistency are assured to maintain.

Categorical Encoding: The function applies integer encoding to transform categorical columns into a numerical format that are used for both determination and demographic columns, e.g. "months_season", "week_or_weekend", "Occupant_ID_in_HH", "region", "number family members" etc. Because LSTM require numerical input.

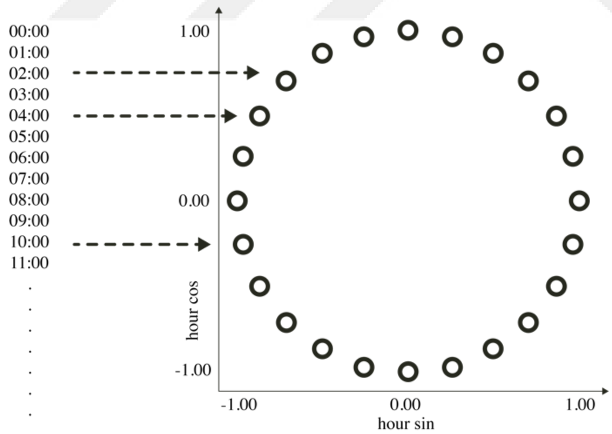

Figure 3.14. Temporal Encoding for continuous features for LSTM

Temporal FE: To capture the cyclic nature of time, sine and cosine transformations are applied to the "hourStart_Activity" and "hourEnd_Activity" features. This encoding effectively preserves the periodic characteristics of these temporal features, as shown in Figure 3.14.

Normalization: The function employs a MinMaxScaler (or StandardScaler) to normalize numerical features. This step is for models that are sensitive to the scale of data, especially for neural networks.

Data Splitting: A stratified splitting approach is applied to divide the data into training, validation, and testing sets. This method ensures that the distributions of the "months_season" and "week_or_weekend" features are consistent across these subsets, which is for unbiased model training and evaluation.

Label Encoding for Target Variable: The "Occupant_Activity" column is transformed using label encoding to prevent any ordinal order for the learning process (Potdar et al., 2017).

Data Reshaping for Sequential Modeling: Finally, the data is reshaped into a format suitable for LSTM models. The input features and target variables are restructured into three-dimensional arrays, compatible with the LSTM's requirement for sequential input, i.e. [batch size, sequence length, feature dimension].

LSTM Model Architecture with Embedding Layers

The proposed model is designed to work with sequential non-sequential datasets. The non-sequential information is inserted in the LSTM architecture with embedding structure thus, the model can perform effectively understand sequences, and embedding layers, simultaneously.

Model Creation:

- Set embedding dimensions for 'months_season', 'week_or_weekend', and other categorical columns.
- Create input and embedding layers for categorical columns; reshape for LSTM.
- Add input layer for continuous features.
- Concatenate embeddings and continuous input.

- Implement Bidirectional LSTM with regularization and batch normalization.
- Add output layer using softmax activation.
- Compile model with optimizer and loss function.

Test Embedding Model:

  - Create input and embedding layers for 'months_season' and 'week_or_weekend'.
  - Reshape embeddings and construct test model.

Train and Evaluate Model:

  - Set up early stopping and model checkpointing.
  - Train model using separated feature sets (categorical and continuous) and target labels.
  - Evaluate model performance on validation and test data.

Evaluate and Save Results:

  - Predict with the model on test data.
  - Convert one-hot encoded predictions to labels.
  - Compile and save actual and predicted labels to a CSV file.

Figure 3.15. Pseudocode of LSTM with Embedding for CENSUS Dataset

Embedding of Non-Temporal Categorical Variables: The model begins with embedding non-temporal categorical variables using dense layers, e.g. “region”, “employment status” etc. This step transforms sparse, high-dimensional categorical data into a dense, lower-dimensional, and continuous space, which is better than traditional encoding methods, e.g., one-hot-encoding.

Embedding of Temporal Categorical Variables: Temporal categorical variables undergo a similar embedding process, i.e., “months_season”, “week_or_weekend”. The embedding of these variables is important for capturing the temporal dynamics of sequential information.

Processing Temporal Data for LSTM: The core of the model consists of the LSTM layers that receives the embedded temporal data along with continuous temporal features. LSTMs are chosen in the model because they are effective for solving the “vanishing gradient” problem. Thus, LSTM can capture long-term dependencies and varying temporal sequences (Noh, 2021; Van Houdt et al., 2020).

Concatenation of LSTM Output with Non-Temporal Data: The output from the LSTM layer is then concatenated with the processed non-temporal data. This concatenation allows the model to integrate insights from both temporal and non-temporal sources.

Final Dense Layers for Classification: Following the concatenation, the combined data is passed through a series of Dense layers. These layers serve as the final step in the model, where the actual prediction is made.

Branching Output Layers: For two target variables, a branching is applied for last layer, i.e. multi-categorical classification for "Occupant_Activity" and binary classification for "Location". This output uses SoftMax activation function that is proper for multi-class classification tasks. It provides a probabilistic interpretation for each category of the "Occupant_Activity" variable. The second output employs a sigmoid activation function, ideal for binary classification tasks.

Batch Normalization (BN) for LSTM

LSTM models are prominently chosen for their ability to capture temporal dependencies in sequential data. The conventional application of LSTM models does not include any batch normalizations instead dropout technique is suggested (Cooijmans et al., 2017). However, BN is preferred in this study for classifying data in 48-row intervals due to independence of intervals, variable scales, and stabilizing learning.

- BN is applied because the independence of intervals is acknowledged. Each 30-minute interval is considered an independent entity with its own distribution of values. The aim is to standardize each feature to have a mean of zero and a standard deviation of one across the batch. Hence, normalizing each interval across the batch is logical because each interval consists of different activity which can easily change from the precedent or subsequent activities.

- BN is applied for the issue of variable scales is addressed by normalizing for the purpose of facilitating the LSTM in learning process. Given that different intervals may exhibit activities occurring at varying frequencies and scales. By providing values that are on a comparable scale, the model can better interpret and learn from the data.
- BN is selected for the stabilization of learning, which is a critical factor. BN is known for its capacity to stabilize and expedite the learning process by mitigating internal covariate shift (Garbin et al., 2020). This aspect is particularly beneficial in scenarios where activities across different intervals exhibit wide variation.

The placement of BN should be placed before the LSTM layer; thus, it can normalize input data without disrupting internal dynamics of the model.
In the case of LSTM models, even though BN is not universally recommended, experimentation can yield beneficial outcomes. In practice, tests have shown that applying batch normalization enhances the performance of the classification task, particularly in scenarios involving the processing of 48 rows per timestep.

- Temporal BN normalizes across all timesteps for a given feature.
- Layer Normalization is normalizing across features for each timestep, and it is preserving the sequential format. It can be implemented in LSTM cells.
- Modified BN applied to LSTM inputs.

The implementation involves a specific process for applying BN after LSTM layers or before the output layers. The shape of the LSTM output is typically `(batch_size, 48, num_features)`. This output is then reshaped to `(-1, 48)` to effectively flatten the features and batch size into a single dimension while keeping the time steps distinct, i.e., 48 rows. At this point, BN is applied to the reshaped tensor for every 48 intervals. BN normalizes the activations of each time step within a mini batch. For each mini batch at time step $\boldsymbol{t}$, the mean $\boldsymbol{\mu_B^t}$ in Eq. (1) and the variance $\boldsymbol{\sigma_B^{t^2}}$ in Eq. (2) are computed as follows:

$$\mu_B^t = \frac{1}{m}\sum_{i=1}^{m} h_i^t \tag{1}$$

$$\sigma_B^{t^2} = \frac{1}{m}\sum_{i=1}^{m}(h_i^t - \mu_B^t)^2 \tag{2}$$

where $\boldsymbol{h_i^t}$ represents the hidden state of instance $\boldsymbol{i}$ at time step $\boldsymbol{t}$, and $\boldsymbol{m}$ is the batch size. The hidden states are then normalized using the formula in Eq. (3):

$$\hat{h}_i^t = \frac{h_i^t - \mu_B^t}{\sqrt{\sigma_B^{t2} - \epsilon}} \tag{3}$$

where ϵ is a small constant added for numerical stability. To ensure the normalized values retain the expressive power required by the model, they are subsequently scaled and shifted using learnable parameters γ^t and β^t, as explained in Eq. (4):

$$\tilde{h}_i^t = \gamma^t \hat{h}_i^t + \beta^t$$

This process is handled by the **nn.BatchNorm1d** function that performs these computations automatically for each feature dimension. After BN, the tensor is reshaped back to its original form for further processing by the next hidden layer or output layer.

Hyperparameter Tuning with Bayesian Optimization (BO)

The selection of appropriate hyperparameter values is crucial for performance optimization of the model. Hyperparameter tuning (HT) involves adjusting learning rates, regularization parameters, and the number of neurons or layers within the network to find an optimal configuration that promotes a higher level of accuracy.

The optimization of the model architecture may involve exploring alternative activation functions, adjusting the depth and breadth of the network, and experimenting with different types of embedding techniques (i.e., sizes). Traditional methods (e.g., grid search or random search) are often computationally expensive.

This study employs the BO for hyperparameter optimization (Nguyen, 2019). The tuning process is applied with using Optuna framework (Akiba et al., 2019). BO is a probabilistic model-based method to globally optimize black-box functions that are expensive to define (Nguyen, 2019). It operates on the principle of constructing a probability model of the objective function and utilizing this model to make informed decisions about which hyperparameters to evaluate for the next iteration.

Implementation Optuna for BO as the Tree-structured Parzen (TPE) algorithm represents a significant advancement over traditional methods. TPE models the search space by constructing two probabilistic models: one for the likelihood of hyperparameters given observed acceptable results and another for hyperparameters given not acceptable results (Bergstra et al., 2011). TPE identifies hyperparameters that have a higher probability of yielding improved outcomes by comparing these two models. Then, it is guiding the search process towards more accurte regions of the hyperparameter space.

BO for LSTM with Embedding Layers

The performance of the model in each trial was assessed based on a predefined metric (e.g. validation loss). This is guiding Optuna in its subsequent search iterations. The validation loss for 'location' column is calculated as the average of binary loss Eq. (1), specifically using BCE With Logit Loss. For 'Occupant_Activity', a multi-categorical loss Eq. (2) is chosen, specifically using Cross Entropy Loss.

$$Loss(x, y) = N * \sum_i (y_i - \log\big(\sigma(x_i)\big) + (1 - y_i) - \log(1 - \sigma(x_i)) \tag{1}$$

$$Loss(x, class) = -\log(\textstyle\sum_j \exp\,(x[j]) \exp\,(x[class])) \tag{2}$$

The process of HT computation begins with Optuna creating an optimization session in which iteratively tests different hyperparameter combinations. In each attempt (i.e. “trial”) a set of settings are tested by running a specific action and then measuring how well it performs. Optuna then updates its internal TPE model based on the outcomes of these trials. Over successive trials, Optuna’s model becomes increasingly adept at predicting which hyperparameters are likely to yield better results. Thus, it is refining the search and converging towards the optimal solution.

One of the important features of Optuna is its ability to “prune” trials. Pruning refers to the early stopping of trials that are unlikely to result in an optimal solution, based on intermediate results. This feature significantly reduces computational cost by preventing the allocation of resources to unpromising trials.

3.4.3 Occupancy Integration: Conversion from TUS Augmented to TUS Energy

This section is about converting TUS augmented to TUS Energy as for implementing occupancy components for building energy modeling, i.e. HHs, OCCs, SCHs, ACTs. This explanation is planned in Figure 3.16 in terms of, which components are related which operational schedule or energy related schedules (ENERs).

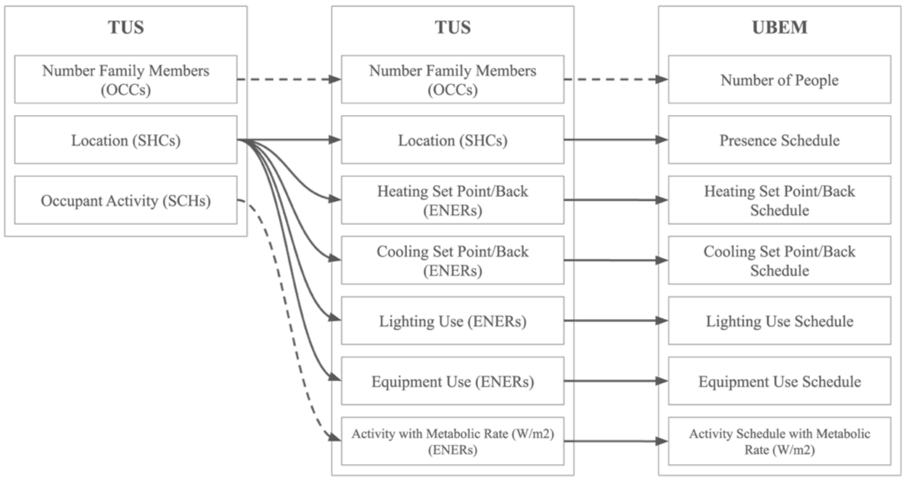

Figure 3.16. From generated occupancy components to UBEM imputs

EnergyPlus (E+) is simulation program is selected as a reference BEM implementation. E+ is a simulation software, and it is chosen as the main tool for this task for clear understanding (Winkelmann et al., 2000). Besides, the implementation is executed using the python framework with eppy and geomeppy libraries for input data file modification (Philip et al., 2011b).

Household Presence Schedule Generation from the TUS dataset

One of the components for the development of the UBEM is the generation of household presence schedules for explaining the existence information of the occupants for during their daily activities. The schedules can be derived from the TUS dataset which is the imputation applied dataset using proposed LSTM with embeddings model. The presence schedules are specifically obtaining the information from the “location” column of each occupant in the household. This column is the 30-min interval representation of each activity that occupants are executed in the residential unit.

Implementation of Household Presence Schedule to UBEM Using E+ Framework

Occupancy presence schedules can be implemented using the objects of “Schedule:Year”, “Schedule:Week:Daily”, and “Schedule:Day:Interval” in E+ for all year presence schedule implementation. As HETUS framework similarly suggest, each season gets represented by one of these weekly schedules and all year is combined with four weekly schedules. Every week schedule is made of three daily schedules for weekdays, Saturdays, and Sundays (or Holidays as it exists in BEM categorization).

The schedule type limits for occupants are either 1 or 0, thus 1 means an occupant is present, and 0 means no one is present. However, for the cumulative version of the presence schedules, household presence schedules are continuous between 0 to 1 with decimals. Thus, the presence schedules are implemented as household schedules.

The generated 12 different daily presence schedules are obtained from the TUS augmented dataset and it is implemented using 'Schedule:Day:Interval” by specifying the weekday categories, i.e. Diary type. Then, according to “months_season” information these daily schedules are merged under the object of “Schedule:Week:Daily”, as weekly representation for each season. Finally, these four different representative schedules are inserted to the “Schedule:Year” for all year representation.

Finally, these intricate yearly schedules are systematically assigned to each thermal zone within the *.idf* files, using specific “People” and “Zone” objects. This methodical integration represents that the occupancy patterns can be accurately reflected in the thermal modeling of each zone.

Conversion From Presence Schedules (SCH) to Operational Schedules (ENER)

The presence schedule of occupancy is the source of the heating set point/back, cooling set point/back, lighting usage, and equipment usage schedules for the simulations. The core of the process is to iteration over combinations of “month_season” and “week_or_weekend” categories because they are representing the different periods of activity daily schedules. Each cumulative presence schedule is created with a similar format in the activity schedules in the TUS dataset, i.e. for each half-hour interval throughout the day. This is achieved by assessing whether the household is occupied during each interval, based on the start and end times of activities of occupants, as well as their locations (i.e. inside or outside the home).

The schedules for heating and cooling adjust the setpoints based on whether the residence is occupied or not for set point/set back selection. Similarly, the usage patterns for lighting and equipment are dynamically configured to reflect the presence or absence of occupants. This conversion ensures that the energy demand modeled in the UBEM is closely aligned with real-life scenarios.

Lastly, density schedules for lighting and equipment energy demand are not evaluated alongside usage schedules. Neither TUS nor CENSUS datasets do not include information on the relation between activity and equipment or lighting usage. This modeling process is a statistical or data related issue. Thus, it is challenging to model density schedules from activities without additional data. For instance, refrigerators are common in households, but televisions are not. Therefore, usage schedules linked to presence schedules are the focus of this study. Each of these inputs reflects how the presence and absence of occupants affect building energy needs.

UBEM implementation of ENER Using E+ framework

All the operational schedules follows the same procedure in terms of E+ time objects, i.e. *'Schedule:Year'*, *'Schedule:Week:Daily',* and *'Schedule:Day:Interval'* as shown in Figure 3.17.

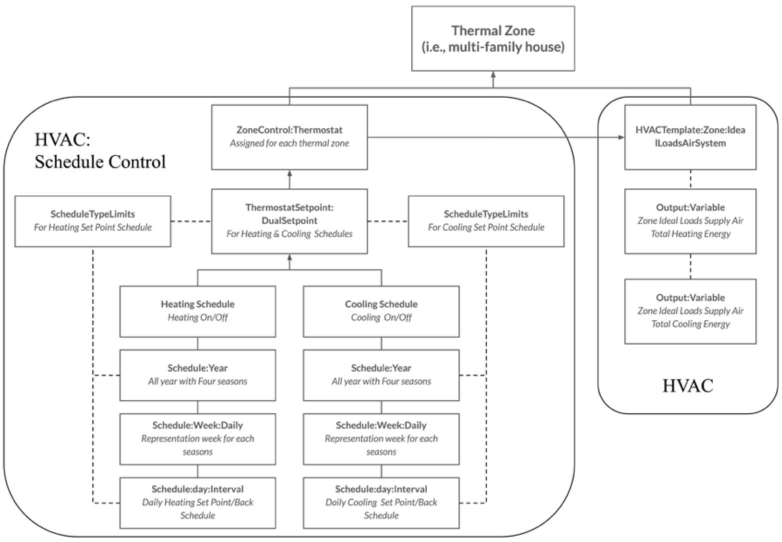

Figure 3.17. Example of Heating set point E+ structures

Four distinct weekly schedules are prepared for the entire year as each season represented by one specific weekly schedule. These schedules consist of three daily types: for weekdays, Saturdays, and Sundays. Schedule type limits are binary, either a set point or a setback, detailed in Table 3.4. None of the standards can completely cover the simultaneous implementation of the cooling and heating systems. Thus, a combination of Italian and USA standards is proposed as the determination of the schedule related values. Besides, an additional randomness is proposed to represent the stochstic dynamics od the occupant behavior by combining two standards values as shown in Table 3.4.

Table 3.4 Set point/back standards for Building Energy Modeling

	Heating Set point (degree)	*Heating Set back (degree)*	*Cooling Set point (degree)*	*Cooling Set back (degree)*
ASHRAE 55, USA (ASHRAE, 2004a)	20/24	15.5/18.3	-	-
ASHRAE 90.1, USA (ASHRAE, 2016)	-	-	23/26	24/26
UNI EN 15251, Italy (Comitato Termotecnico Italiano, 2011)	19/22	16/18	-	-
UNI EN 11439, Italy (Comitato Termotecnico Italiano, 2013)	-	-	26	28
The combination of the standards	20	18	26	28
If to add variation	random (19,22)	random (16,18)	random (23,26)	random (24,26)

Before implementing the heating and cooling operational schedules, the operational periods of the two conditioning systems should be determined. Because the periods of the two conditioning systems are different, for ins., it is not possible to open heating system during summer or unnecessary to operate thus, periods should be defined by covering all year period.

Table 3.5 Operation Periods of the Conditioning Systems

	Cooling Off Heating On	*Cooling Off Heating On*	*Cooling On Heating Off*	*Cooling On Heating Off*	*Cooling On Heating Off*	*Cooling Off Heating On*	*Cooling Off Heating On*
Seasons	*Winter*	*Spring*	*Spring*	*Summer*	*Autumn*	*Autumn*	*Winter*
Start Month	1	4	4	6	9	10	12
Start Day	1	1	16	1	1	16	1
End Month	2	4	5	8	10	11	12
End Day	28	15	31	31	15	30	31

Table 3.5 presents the operation periods of the cooling and heating systems. This implementation is the simplest usage for periodical determination besides from that heating or cooling degree days can be extracted from the weather files of the E+ and the periodical divisions can be executed (Büyükalaca et al., 2001; Mehregan et al., 2022). These periodical divisions are implemented using the *'Schedule:Week:Daily'* E+ objects. For simpler explanation, either mechanical or natural ventilations are excluded from the periodical table.

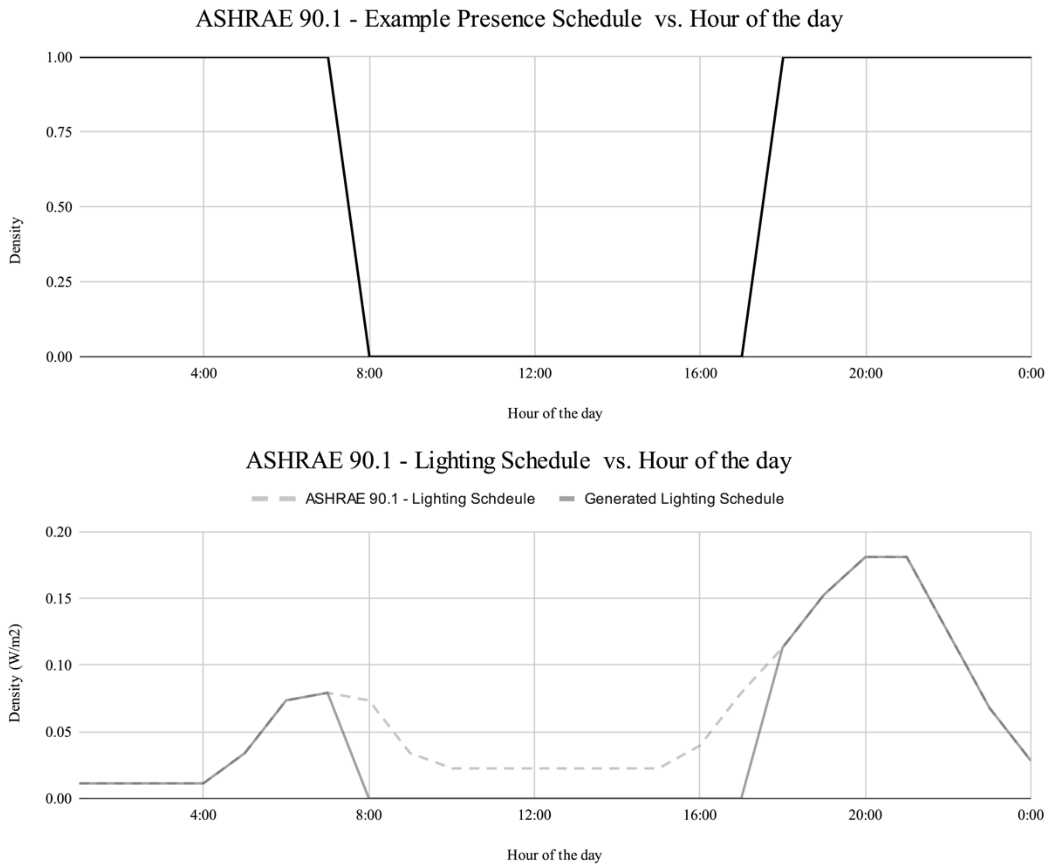

Figure 3.18. Generation of Ligthing Use schedules using Presence Schedules

Implementation of lighting schedules is similar to that of the presence schedule. Unlike the set point schedules, these schedules are continuous and highly dependent on the number of people and their combined presence schedule. Since there are no international standards for usage schedules, it is proposed to use default ASHRAE standards as a base schedule (ASHRAE, 2013; DOE, 2013). This schedule becomes active when someone is present in the residential unit and non-active when the unit is unoccupied. The baseline schedule is selected from the ASHRAE 90.1 standards as lighting schedule template for "midrise lighting level". Besides the power density is selected between 6-10 w/m^2 according to similar Italian context BEM research (Bianco Mauthe Degerfeld et al., 2023). Figure 3.18 presents the multiplication of the presence schedule with the template lighting use schedule from ASHRAE to generated example lighting use schedule for each occupant. Each occupant has different presence schedules for each daily activity thus this generated lighting use schedules has the variation.

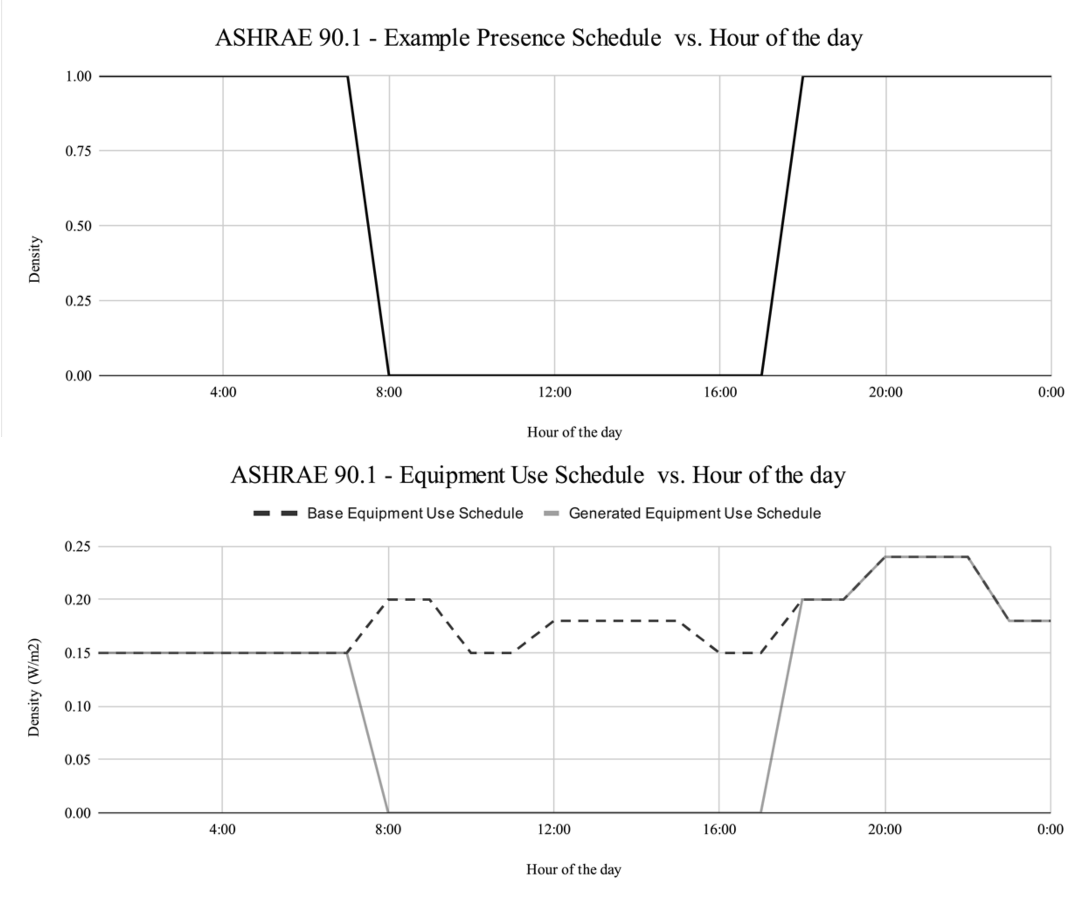

Figure 3.19. Generation of Equipment Use schedules using Presence Schedules

The scheduling of equipment usage mirrors that of lighting schedules, as both are dependent on the occupancy patterns of the household. Consequently, a base schedule for equipment use is necessary because there is a lack of a standard equipment use schedule for the Italian region. Thus, the foundation for this base equipment uses schedule is sourced from the research referenced (FEDERICA ZAGARELLA, 2019). Similar computation of lighting use schedules, the base schedule is multiplied with the presence schedule for each 30-minute interval. This results in distinct equipment use schedules for each occupant as shown in Figure 3.19. These schedules are then incorporated into the 'ElectricEquipment' object in EnergyPlus, using 'Schedule:Year', 'Schedule:Week:Daily', and 'Schedule:Day:Interval'. However, equipment schedule modeling could consider occupant density and equipment type, but this necessitates additional data collection. Therefore, equipment use schedules are exclusively generated for each occupant, i.e. for 12 different daily presence schedules.

Household Activity Schedule Generation from TUS

The conversion process from TUS sequential schedules to UBEM inputs extends beyond the generation of presence schedules to encompass the integration of metabolic rate-based activity schedules from the classified activity schedules. The conversion is the combination is TUS augmented and metabolic rate conversion chart which is derived from the Compendium of Physical Activities (2011). The content is the metabolic rate (W/m²) equivalent for each activity (Ainsworth et al., 2011).

The conversion starts by from the metabolic rate conversion of each occupant then these metabolic rate values summarized under the selected household. The conversion aggregates the metabolic rates ($\boldsymbol{M}$) of all individual activities ($\boldsymbol{i}$) carried out by the occupants present in the home at a given hour ($\boldsymbol{M_i}$) while generating metabolic rate equivalents of activities for the entire household in Eq. (1). Then, aggregated value is divided by the number of people ($\boldsymbol{P}$) present in the house at that time. This refinement ensures that the cumulative activity schedules generated are a true representation of the household's overall metabolic energy expenditure at any given hour by considering for the presence and activity level of each occupant. The genuine aspect of this approach is its capability to sum these metabolic rates and average them per family member within each household.

$$average\ MET\ per\ person = \frac{\sum_{i=1}^{n} Mi}{P} \qquad (1)$$

This metabolic rate data is then incorporated into E+ simulations through the 'Activity Level Schedule Name' in the PEOPLE object of the IDF files. The schedule consists of also the zone name, people density, and the activity level schedule.

Conversion From Activity Schedules to Metabolic-rate Activity Schedules

Metabolic-rate Activity Schedules increase building energy simulations (BES) accuracy in terms of thermal comfort parameters (S. H. Hong et al., 2018; Sood et al., 2023). Internal heat gains are correlated to occupant presence, and they can be modeled considering the time during which the dwelling is occupied with the number of occupants in the dwelling and their performed activities (Buttitta et al., 2019). In technical explanation, internal heat gain computation is linked to metabolic rate of the occupants with their presence and activity information i.e, 'Occupant_Activity'.

CODE	METS	MAJOR HEADING	SPECIFIC ACTIVITIES
05057	*3.0*	home activities	cooking Indian bread on an outside stove
05060	2.3	home activities	food shopping with or without a grocery cart, standing or walking
05065	*2.3*	home activities	non-food shopping, with or without a cart, standing or walking
05070	1.8	home activities	ironing
05080	1.3	home activities	knitting, sewing, light effort, wrapping presents, sitting
05082	2.8	home activities	sewing with a machine
05090	2.0	home activities	laundry, fold or hang clothes, put clothes in washer or dryer, packing suitcase, washing clothes by hand, implied standing, light effort
05092	4.0	home activities	laundry, hanging wash, washing clothes by hand, moderate effort
05095	*2.3*	home activities	laundry, putting away clothes, gathering clothes to pack, putting away laundry, implied walking
05100	3.3	home activities	making bed, changing linens
05110	5.0	home activities	maple syruping/sugar bushing (including carrying buckets, carrying wood)
05120	5.8	home activities	moving furniture, household items, carrying boxes
05121	5.0	home activities	moving, lifting light loads
05125	4.8	home activities	organizing room
05130	3.5	home activities	scrubbing floors, on hands and knees, scrubbing bathroom, bathtub, moderate effort
05131	2.0	home activities	scrubbing floors, on hands and knees, scrubbing bathroom, bathtub, light effort
05132	6.5	home activities	scrubbing floors, on hands and knees, scrubbing bathroom, bathtub, vigorous effort
05140	*4.0*	home activities	sweeping garage, sidewalk or outside of house
05146	*3.5*	home activities	standing, packing/unpacking boxes, occasional lifting of lightweight household items, loading or unloading items in car, moderate effort
05147	*3.0*	home activities	implied walking, putting away household items, moderate effort

Figure 3.20. Metabolic Rate of Human Activities

The metabolic-rate conversion resource of 2011 Adult Compendium of Physical Activities is selected over the standard ASHRAE 55 due to its extensive range of activities with varied metabolic rates (ASHRAE, 2004b). It lists activities from Indian cooking bread on an outside stove (3.0 W/m²) to sewing with a machine (2.8 W/m²) as shown in Figure 3.20.

The implementation of these schedules for BES involves the usage of several strategic E+ objects which are used for other operation schedules, e.g. lighting use etc. Consequently, *'Schedule:Day:List'*, *'Schedule:Week:Daily'*, and *'Schedule:Year'* are employed to map out the occupant activities. Daily schedules are assigned in 30-minute intervals. The metabolic rate conversions are applied strictly to the activities themselve. There is no location-based conversion implemented. This approach ensures that the energy impact is directly correlated with the specific activities without adding any complexity of varying location factors. For instance, sleeping is usually executed in the bedroom or cooking in the kitchen however, there is a need for zone-based residential unit modeling. The simplicity of this methodology aids in maintaining clarity in the simulation process, while still capturing the essential dynamics of occupant behavior in relation to building energy demand.

Every daily activity listed in the Time Use Survey (TUS) is carefully matched with an equivalent activity from the Compendium, 2011. This matching is essential to accurately capture the metabolic rates for each activity. Activities, e.g. "Main Job" or "Cooking, Washing and Tidying Up Dishes" from TUS are paired with similar activities in the Compendium. The metabolic rate values are then taken from the Compendium, e.g. Activity Level W/m². These values are converted following the standards set by the ASHRAE 55 metabolic rate conversion table. This table provides baseline values like BaseValue: Activity Level w/ Person EnergyPlus Schedule Value and BaseValue: Activity Level W/m². The detailed conversion process ensures that the simulation accurately represents the energy demands of daily activities in the built environment.

Metabolic Rate & Taylor Code and Compendium, 2011

The Taylor Code is a classification framework which is systematically assigning codes to a range of human activities to denote their metabolic intensity (Heydenreich et al., 2019). Each code is indicative of the activity's energy demand for facilitating the estimation of individual energy expenditures. This framework is used in ergonomics, workplace environment design, and occupational health studies. In conclusion, the 2011 Compendium of Physical Activities adopts the Taylor Code standard, using it as a foundational reference for categorizing activities based on their metabolic rates.

Assignation of Household Information for Residential Units

TUS and CENSUS datasets are both collected by ISTAT, however they cannot be merged into a single dataset due to mismatching identification columns for households and residential units, as well as different categorical definitions. For example, while the TUS dataset includes a column for family member counts to six, the CENSUS dataset extends this to seven. Additionally, the age category classifications vary between the two datasets. Despite these differences, the CENSUS dataset is valuable because it is a comprehensive overview of Italian households and links residential unit details with household data. CENSUS provides distribution data that can be used when assigning households for BEM simulations. This is important because the TUS dataset lacks information on residential physical properties, i.e. floor area and room count.

The CENSUS dataset provides a detailed view of the residential unit floor areas. They are ranging from 20 to 140 square meters as multi-house structures. The graphical representation of these data underscores the variation in living spaces, which directly correlates to differing residential dynamics. Consequently, the household size column is both exist in TUS and CENSUS datasets thus, the assignation of the household can be executed using the household size according to the floor area of the residential units.

CHAPTER 4

RESULTS

This chapter presents the study's results on occupancy modeling and its implications in UBEM. The chapter covers data preprocessing methods applied to the TUS and CENSUS sub-datasets, trends observed in total activity numbers, and the impact of data augmentation techniques. The chapter focuses on the presentation of the accuracy and computation performance of the proposed method (i.e., End-to-End Multi-Task Deep Learning Model) for occupancy modeling under UBEM and its accuracy and learning performance comparison with the alternative methods that are modeled by researchers from the same field.

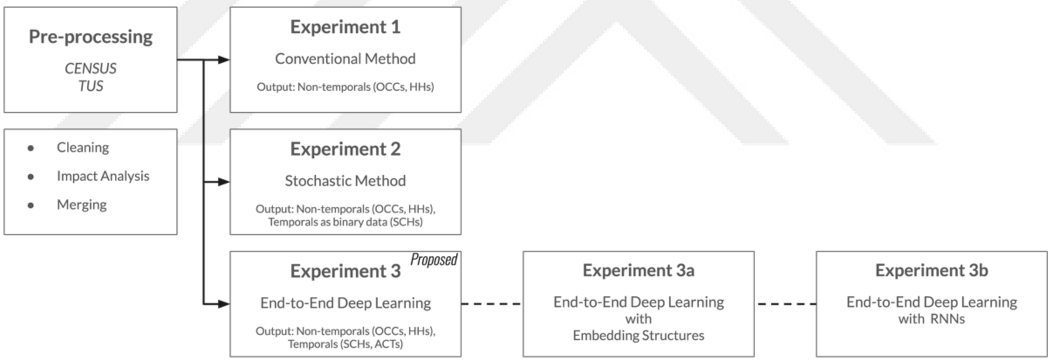

Figure 4.1. Experimental Setup Chart of All Experiments

4.1 Before Experiments: Pre-processing of Dataset

The goal of this sections is to model the original data sets into actionable insights, ensuring their accuracy and relevance for energy modeling purposes using cleaning and transformation techniques. The focus is on selecting suitable data content from two primary sources i.e., TUS dataset and CENSUS dataset.

4.1.1 Preprocessing for Census Housing Sub-Dataset, 2011

The dataset contains information on the location of households, including physical attributes of residences and air conditioning features. Data preprocessing involved several key steps to refine the dataset for occupancy generation steps, as shown in Figure 4.2. Columns is unrelated with UBEM were eliminated, and ineffective or singular pieces of information were consolidated. Subsequently, manual imputation was conducted, guided by the interrelationships among columns, to address any deficiencies in the data. An analysis of feature importance was performed, specifically focusing on columns related to household and residential characteristics to inform further research and modeling efforts. Most of the columns are only preserved until the merging the CENSUS housing and CENSUS occupant, then only residential floor area and number of family members columns are used for subsequent steps.

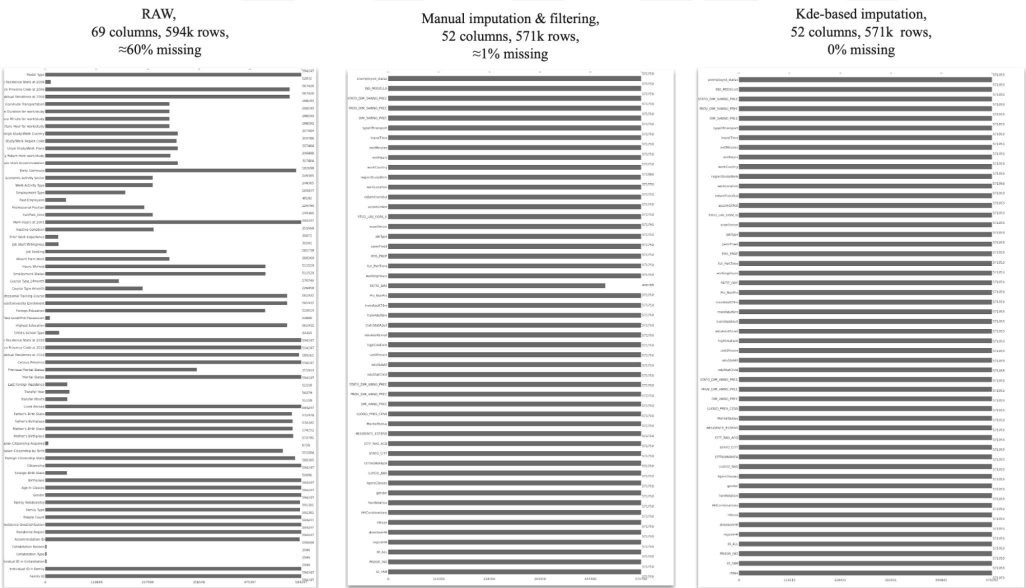

Figure 4.2. The visual proof of the need for imputation, Census Housing

4.1.2 Preprocessing for TUS Individual Sub-Dataset, 2011

Unlike the CENSUS dataset, the TUS subset provides a detailed sequential representation of personal characteristics of household members, including their movements related to work or education. The current analysis excludes considerations of household and occupant demographics which is non-sequential dataset. However, many columns are also with missing rows.

The dataset is organized into specific categories with defined contents. The identification category (IDs) includes Household ID and Occupant ID within the household. Household characteristics (HHs) encompass variables as the number of family members, family typology, and presence of domestic help, as well as locational data, e.g. administrative and orientational region.

Occupant characteristics (OCCs) include gender, age, marital status, childcare, life satisfaction, educational attainment (including type, degree, and diplomas), and employment details (covering job status, sector, work conditions, hours, and commuting details including modes of transportation). However, only the most important features are retained for occupancy modeling, as shown in Figure 4.3.

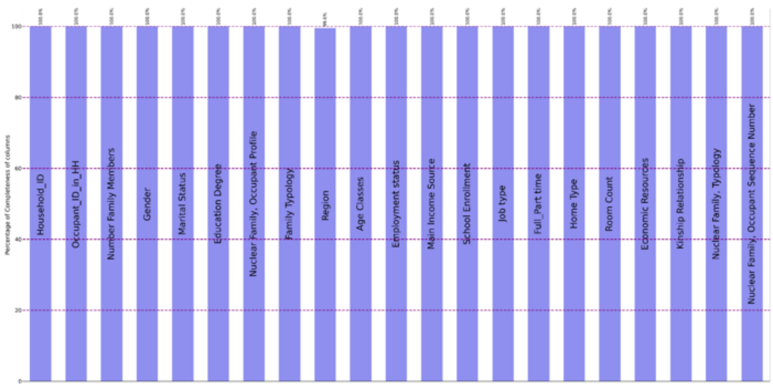

Figure 4.3. Missingness ratio of features, after preprocessing, TUS Individual

The dataset has a missingness rate of approximately 59.23% across its 267 columns and 44,000 rows. Due to the extent of the gaps, manual restitution is not feasible, and both manual and algorithmic imputations are necessary. The preprocessing procedure involved renaming column headers for clarity, removing redundant columns, consolidating columns with negligible or singular data, and implementing manual and algorithmic imputation methods based on the interrelationships between columns. This included iterative imputation and KDE-based distribution imputation. During the preprocessing stage, it was crucial to evaluate the significance of features, especially those related to household characteristics and commuting patterns of the occupants.

4.1.3 Preprocessing for TUS Daily Sub-Dataset, 2011

Each row in the original dataset represents an activity performed. These activities can belong to any member of any household. There are household and occupant *IDs* showing who owns the activities, as well as information about who is accompanying the activity or not. As mentioned above, a daily representation of a person is described, and for this, both the order of activity and the start and end times are specified. Any of the *HHs*, *OCCs* and *RESs* group parameters found in the TUS individual or CENSUS are not available in the TUS daily data set. For this reason, the *HHs*, *OCCs* and *RESs* group parameters of the person who performed the activity were combined with the TUS individual data using *IDs* columns.

Table 4.1 shows the missingness of the TUS daily dataset in terms of temporal and operational status. 1-day activity data of each household and its members in the TUS individual dataset is included in the TUS daily dataset. On the other hand, although these activities differ according to the season, part of the week and working status, everyone has only one daily activity information. From a positive point of view, individuals who are like each other in terms of employment status, age, education, and gender have daily activity information for different days in terms of time. In this way, the whole data set was completed by combining the days of similar occupants.

Table 4.1 TUS daily dataset, missingness in terms of Employment Status

Temporal resolution		*Employment Status, categories*						
season	*TUS_daily*	*Child in home*	*Employed*	*Unemployed*	*Retired*	*Student*	*House person*	*Unable to work*
winter	Weekday(s)	exist	NaN	NaN	NaN	NaN	NaN	NaN
spring	saturday	NaN	exist	NaN	NaN	NaN	NaN	exist
spring	sunday	NaN	NaN	exist	NaN	NaN	NaN	NaN
summer	Weekday(s)	NaN	NaN	NaN	NaN	exist	NaN	exist
summer	saturday	NaN	exist	NaN	NaN	NaN	NaN	NaN
summer	sunday	NaN	NaN	NaN	NaN	NaN	NaN	NaN
autumn	Weekday(s)	NaN	exist	NaN	NaN	exist	NaN	NaN
autumn	saturday	NaN	NaN	NaN	NaN	NaN	NaN	NaN
autumn	sunday	NaN	NaN	NaN	NaN	NaN	exist	NaN

The TUS subset and TUS daily datasets were integrated by utilizing the "HH ID" and "Occupant ID in HH" columns. Additional attributes, such as age classes, employment status, and gender, were incorporated from the TUS individual. The TUS daily dataset was enriched by this process, resulting in the creation of the TUS temp.

Table 4.2 TUS temporal, content

Age classes	Employment status	gender	Adm. region	region	Months _season	Week_or_ weekend	location	Occupant _activity	Starting hour	Finishing hour
5	1	2	5	2	1	2	1	1	4	7
5	1	2	5	2	1	2	1	2	7	8

Next, the dataset was clustered using the k-modes clustering algorithm to categorize each household member into clusters. The algorithm utilized the newly added columns along with the "Occupant ID in HH" to facilitate the clustering. The classification tags resulting from the clustering exercise were used to fill in all missing data related to the season and week segment for each user. This imputation process increased the dataset’s volume from 1 million records to over 10 million records, providing a significant increase in the breadth and detail of the data available for analysis.

4.1.4 Merging for Temporal Features of TUS Sub-Datasets

The methodology for merging the TUS sub-datasets is augmenting the occupancy demographics with sequential information, i.e. activity, presence, and location. The methodology under examination places importance on the process of data integration, especially when it comes to combining datasets with different temporal characteristics. This involves a critical step of merging two distinct datasets, one with non-temporal features (i.e. TUS individual as non-sequential) and the other with temporal attributes (i.e. TUS daily as sequential), in a seamless manner. The approach starts by importing the datasets into the working environment using a widely recognized data manipulation library known for its robustness and reliability in handling large datasets, i.e. pandas-python library.

Temporal resolution	Employment Status, categories*						
TUS_daily**	Child in home	Employed	Unemployed	Retired	Student	House person	Unable to work
Weekday(s)	x		x				x
saturday		x			x		
sunday				x		x	

TUS_daily

Household ID	Occupant ID in HH	Region	Room Count	Number Family Members	Age Classes	Employment Status
1	1	0	1	2	1	1
1	1	1	1	2	1	2

TUS_indiv

Figure 4.4. Merging process without any augmentation for missing schedules

As shown in Figure 4.4, the integration process involves aligning and merging the datasets through a carefully executed join operation governed by specific identifiers, i.e. “Household_ID” as the identification of the household and “OCC_ID_in_HH” as the identification of the order number of the occupant in the family. The significance of this step is its ability to preserve the integrity and usefulness of the datasets, which is essential for subsequent analysis.

4.1.5 Preprocessing for TUS Temporal

The process consists of activity filtering to retain essential details and enhancing resolution by interpolating data to shift from one-hour to 30-minute intervals. This process acknowledges that activities vary in duration; some may extend beyond an hour, while others are shorter, e.g. sleeping is longest mostly. For example, activity schedules of occupants can vary significantly, with one having 80 rows of activities and another just 15, indicating longer-duration activities for the latter.

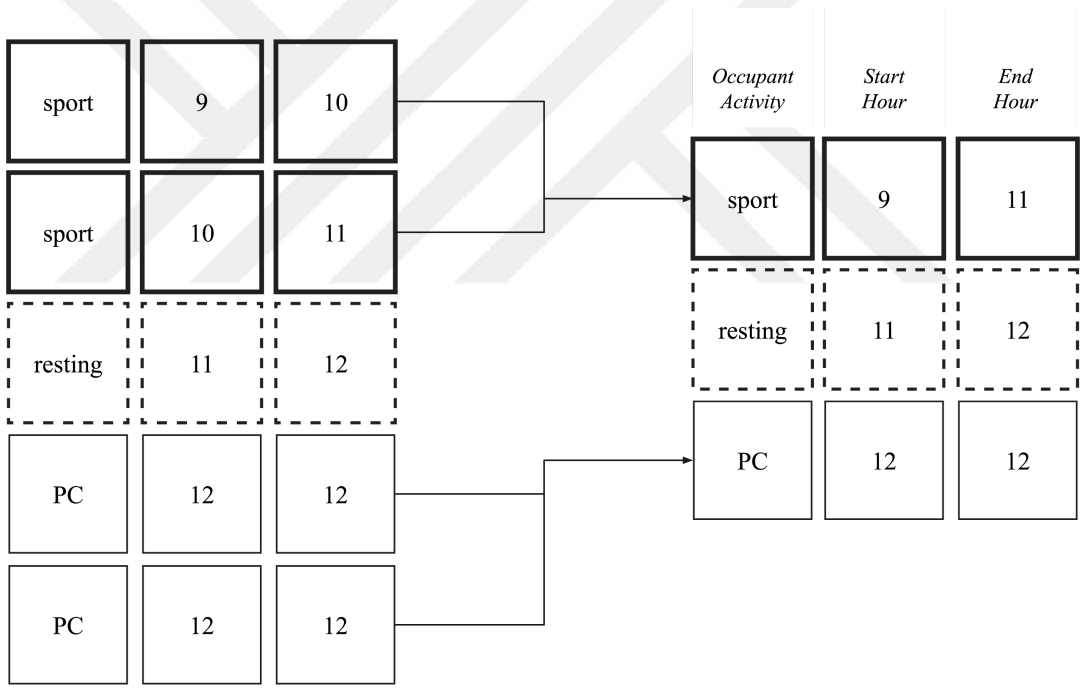

Figure 4.5. Equalizing the length of all daily schedules

Since LSTM models require input in the format of [samples, timesteps, features] thus, padding is essential for constant “timesteps” number. This results as aiding in the handling of padding sequences for LSTM model and standardizing input data to 48-rows per daily activity schedules i.e. 30-min interval per activity, as shown in

Figure 4.5. The dataset, with its refined temporal granularity and structured sequence of household ("Household_ID") and occupant ("Occupant_ID_in_HH") identifiers. After this process, the dataset is optimized for analyzing residential occupancy schedules, capturing the intricate patterns necessary for accurate modeling.

Impact Analysis for Demographic Features

An impact analysis was conducted to assess the influence of various demographic characteristics of occupants on activity and location variables. This analysis focused on examining how selected demographic features varied according to location and activity datasets, e.g. occupant age, household size, etc. The objective was to understand how these demographic factors correlate with and potentially impact occupancy behaviors and trends in different residential settings. As the datasets are temporal, to execute the impact analysis, feature engineering approach is conducted to convert temporal data format to non-temporal data format.

Target variables are sequential, i.e. "occupant_Activity", "location", on the contrary, occupant demographics are non-sequential, i.e. "Region", "Number Family Members", "Family Typology", "Age Classes", "Employment status", "Gender", "Education Degree".

Feature Engineering (FE) of Sequential data for Impact Analysis

The calculation of FE is important for explaining the relationship between occupant demographics and the daily presence and activity routines. Complex behaviors can be explained as interpretable metrics by converting these features with the help of FE. Here are the derived columns:

- Aggregative Features: "distinct activities", "most frequent activity" "least frequent activity"
- Temporal Patterns: "Start_of_day", "end_of_day", "Max consecutive"
- Changing Points: "Activity changes"
- Transformation Techniques: "seasonal", "Activity entropy"

- Location features: “Total_Hours_Outside”, “Total_Hours_Inside”

Aggregative features as “distinct activities” provide a count of the unique activities an occupant engages in, offering a measure of diversity in daily routines. The “most frequent activity” and “least frequent activity” reflect the predominant and least common behaviors. It reveals the occupant’s routine preferences and outlier activities, i.e. walk with the dog. Temporal patterns are captured through “Start_of_day” and “End_of_day”. The features are converted by placing limits for the daily time boundaries of active occupancy. On the other hand, “Max consecutive” quantifies the duration of sustained engagement in a single activity, it can be evaluated as a proxy for activity consistency. The “Activity changes” metric enriches this narrative by enumerating the frequency of transitions between activities. It can be seen as an indicative metric for the occupant’s daily rhythm variability.

Transformation techniques as “seasonal”, it is a decomposition process for the separate the occupant activity data into trend, seasonality, and residual components. This feature directly unravels the underlying structure of time-series data. “Activity entropy” goes a step further and it is providing a statistical measure of unpredictability in activity patterns by distinguishing between random and structured activities. Lastly, location-based features “Total_Hours_Outside” and “Total_Hours_Inside” aggregate the duration spent in different spatial contexts. It can be basically seen as the occupant’s interaction with their living environment. These engineered features, derived from data manipulation techniques, enable a more sophisticated analysis of how demographic factors impact residential energy consumption, serving as a cornerstone for the development of predictive models for occupancy modeling.

Composite Score for Impact Analysis

The conception of a composite score is a synthesis of multiple behavioral indicators which are derived with FE. This score is derived from a two statistical method by combining Principal Component Analysis (PCA) for continuous variables and Multiple Correspondence Analysis (MCA) for categorical variables. Continuous features are “Activity Entropy” and “seasonal”. To ensure all data points are directly comparable, they are normalized. Then, a technique called Principal Component Analysis (PCA) is used to identify the single most important underlying factor that captures most of the data’s variation. At the same time, categorical features are processed through MCA. These include a wide range of occupant activities and temporal patterns.

The original features are replaced within the dataset with this composite score. This score then serves as the foundation for further analysis. It provides a pivotal reference point for assessing the influence of demographic attributes on occupancy patterns, thereby enhancing the predictive robustness and interpretative clarity.

Permutation Feature Importance (PFI)

PFI is employed to analyze the impact of occupant demographics on activity and presence schedules (Altmann et al., 2010). This is a robust model-agnostic technique. PFI measures the decrease in a model’s performance when each feature’s values are shuffled randomly for disrupting the association with the target outcome. A lower score on performance metrics for accuracy or R-squared suggests a decrease in the feature’s ability to predict the outcome. The operational steps involve training the model to establish a baseline performance, shuffling one feature across the validation set, and recording the performance deviation. The magnitude of deviation reflects the feature’s importance. Thus, a greater decrease denotes a more significant role in determining occupancy patterns.

Results of Impact Analysis

Figure 4.6 presented illustrates the relative significance of various demographic variables in predicting the composite score. These findings emerge from a rigorous data-driven approach, aimed at enhancing the precision of occupancy modeling.

"Age Classes" is the most influential demographic factor, strongly correlating with daily activity and location patterns. Following this, "Gender" and "Employment Status" are significant predictors that are indicating their contribution to occupancy variations. "Number of Family Members" and "Family Typology" hold moderate importance which are affecting occupancy schedules.

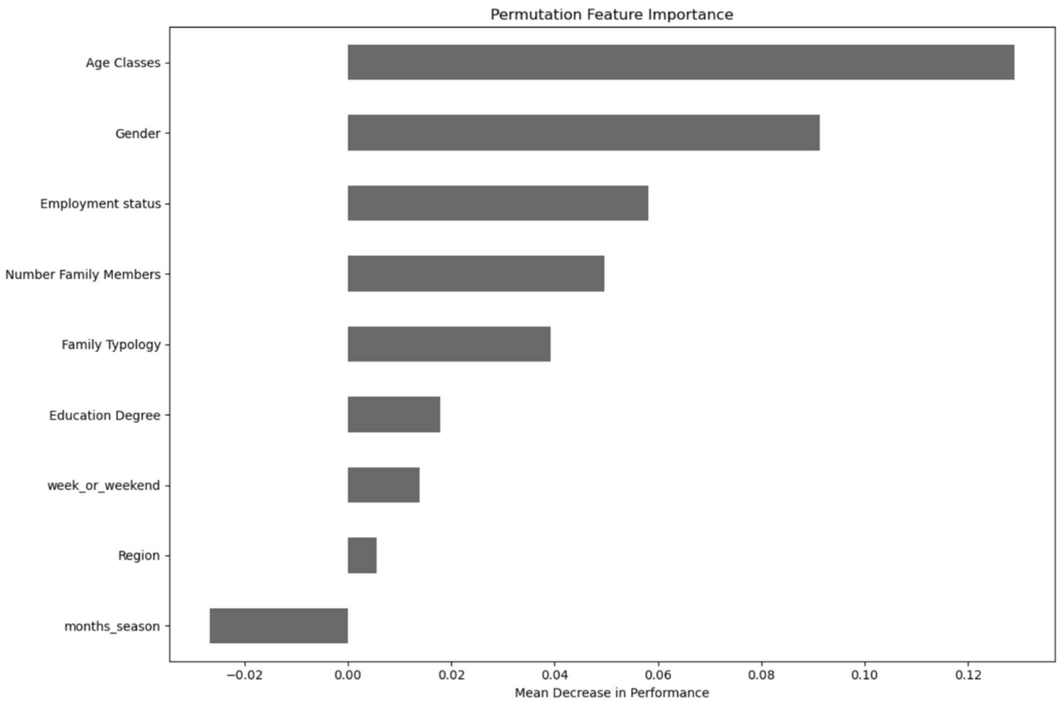

Figure 4.6. Permutation Feature Importance

Finally, "Education Degree" has lower importance, while "week_or_weekend" and "Region" show minimal influence compared to demographic variables.

4.2 Experimental Setup: Occupancy Modeling

In this study, End-to-End Multi-Task Deep Learning Model is proposed for realistic occupancy modeling under UBEM simulation accuracy. However, it is necessary to compare the model different approaches to present how the performance of the occupancy modeling has developed. Thus, this section presents a comprehensive comparison of the proposed model against a variety of other models to evaluate its effectiveness in occupancy generation. The comparison spans several categories as each providing unique insights. It is aimed to clearly explain how well the proposed model performs.

Firstly, the proposed model is compared with Conventional Method of Clustering with Feature Engineering. This method, though not parallel in its approach to the LSTM model, it is proposed from different researchers about occupancy modeling and the method offers a valuable perspective as it encapsulates the visual representation of clusters in occupancy modeling.

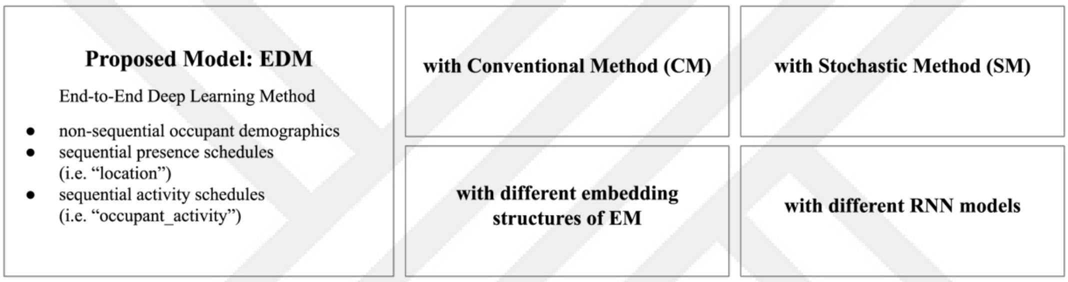

Figure 4.7. Comparisons of the proposed method based on literature review, previous methods

Secondly, the model is compared with stochastic-based occupancy generation models, including Markov Chain models. Different from the Markov model, proposed LSTM model consist of embedding layers that incorporate week and season features along with occupant demographics. In the existing literature, numerous applications of the Markov process for occupancy modeling focus predominantly on presence schedules, without addressing multi-categorical sequential activity classification. Contrasting this, the current study experiments with the Markov process for activity classification, aiming to demonstrate the capability of the LSTM model in handling this more complex task. This approach highlights the advanced potential of LSTM models in the realistic occupancy modeling.

Thirdly, the analysis extends to different embedding structures, contrasting the proposed LSTM model with variations that use only week and season features, as well as those without embedding layers. This comparison helps to identify the impact of the occupant demographics for sequential activity and presence schedule classifications. It is positive that embedding structure is advantageous to combine non-sequential data format of demographics with sequential format of daily schedules.

Lastly, a comparison is conducted with another recurrent neural network (RNN) models conducted. Both models also integrate week and season features with occupant demographics by using embedding structure.

4.3 Experiment 1: Conventional Method (CM): Machine Learning with FE

This section explores the integration of FE for ML as a conventional method and the performance comparison with the EM (Shetty & Singh, 2021). It focuses on enhancing model performance through the strategic manipulation of sequential data features as seen in Figure 4.8.

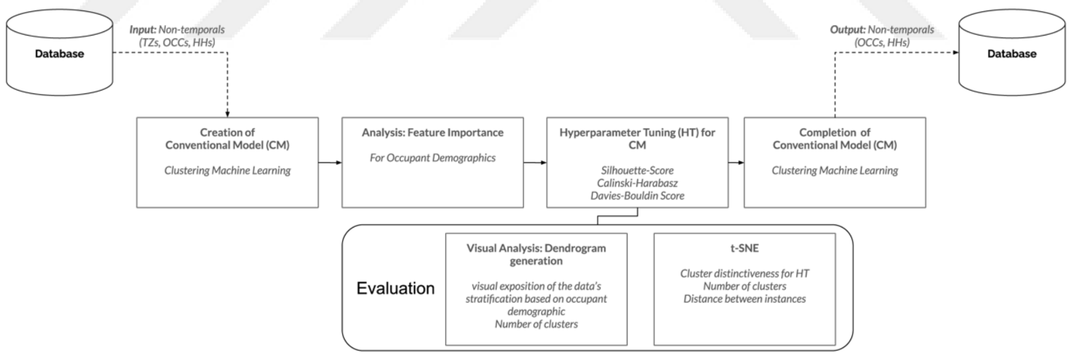

Figure 4.8. Application Steps of Conventional Method

Comparison of various embedding structures and recurrent neural networks, the evaluation was grounded in quantitative metrics, i.e. accuracy and loss for training and validation datasets. These kinds of direct comparisons are reasonable when assessing models of a similar nature and objective i.e. sequential deep learning.

However, a paradigm shift is observed when juxtaposing the LSTM (i.e. end-to-end deep learning model) and clustering algorithm (i.e. ML with FE).

The significant differences in the structure and purpose of the models make it impractical to compare them using standard measures like accuracy and loss. The LSTM model is evaluated on its ability to generate temporal classifications. On the other hand, the clustering approach segments occupant profiles and assigns existing activity schedules to impute missing daily activities. It does not inherently generate new data but rather reallocates existing information. Therefore, a comparison based on similarity performance of the generated versus actual daily schedules is not a viable measure for these methodologies. Consequently, another chapter is dedicated to explaining the deficiencies of the clustering approach for realistic occupancy modeling.

4.3.1 Hyper-parameter Tuning (HT) of CM

In the first step as a initial phase of HT, a comprehensive exploration of agglomerative hierarchical clustering was executed applying variations for number of clusters (n_clusters), the method of linkage, and the metric for distance calculation. During the process, the quality of each clustering configuration was evaluated using three selected metrics, i.e. the Silhouette score, the Calinski-Harabasz score, and the Davies-Bouldin index (Januzaj et al., 2023; Petrovic, 2006; Sudarno et al., 2022). Thus, the process of determining the ideal range for the number of clusters (n_clusters) was guided by configurations that regularly produced high Silhouette and Calinski-Harabasz scores, coupled with low Davies-Bouldin scores (Saitta et al., 2008).

- The Silhouette score, with its range extending from -1 to 1, a higher score means better separation.
- The Calinski-Harabasz score, a higher score means indicative of more clearly delineated clusters.

- The Davies-Bouldin score, a lower score is preferable.

During the exploration of optimal clustering configurations, several key observations and determinations were made. Initially, a notable constraint was encountered with the “ward” linkage method, which showed computational errors when paired with non-Euclidean distances like Manhattan and Mahala Nobis. Thus, these combinations are excluded from the tuning computation. Additionally, Mahala Nobis metric requires specifying extra parameters and was incompatible with the “single” linkage method, thus, leading to errors and it is excluded from the hyper-parameter space.

As a result, the combination of the Ward linkage method with the Euclidean distance metric emerged as the most effective as yielding superior results. A notable trend was the improvement in the Silhouette score as the number of clusters increased, especially with the combinations of the Ward linkage and Euclidean metric. However, this trend had to be carefully balanced with the observations in the Calinski-Harabasz and Davies-Bouldin scores to achieve an optimal clustering outcome.

Hyper-parameter configurations for the selection of optimal number of clusters (n_clusters) generated high Silhouette and Calinski-Harabasz scores, and low Davies-Bouldin scores. A trend was noted where the Silhouette score improved with an increase in cluster numbers. However, this improvement needed to be balanced against the trends in the Calinski-Harabasz and Davies-Bouldin scores to achieve the robust evaluation of the clustering configurations.

For the initial phase, using the Ward linkage method and Euclidean metric, configurations with 40 and 50 clusters proved particularly effective. High Silhouette scores indicated strong intra-cluster cohesion and clear inter-cluster distinction. The Calinski-Harabasz score peaked at lower cluster counts for 10 and 20. It decreased with larger cluster numbers, suggesting more defined clusters at these lower counts. Additionally, lower Davies-Bouldin scores were often seen with the Ward linkage and Euclidean metric. This indicated effective clustering. In conclusion, the analysis suggests an optimal range for the number of clusters (n_clusters) to be within 40-50. This range is chosen based on high Silhouette scores, showing well-defined cluster separation. There is also a balance in Calinski-Harabasz and Davies-Bouldin scores. This balance suggests optimal cluster density and minimal variation within clusters.

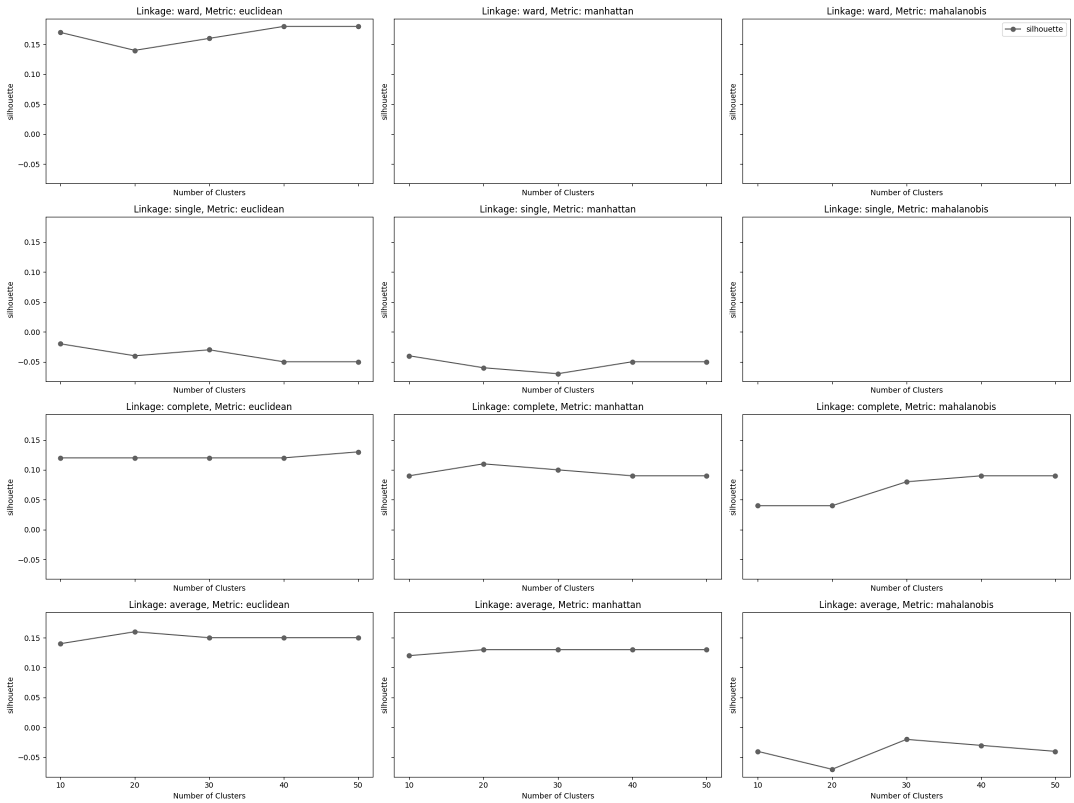

Figure 4.9. Tuning progress for the silhouette scores, rest of the plots are in appendix

In the second step as a refinement phase of HT, a detailed evaluation was carried out to find the best setup for agglomerative hierarchical clustering with narrower range values for hyperparameter tuning. This was done within a specific range of cluster numbers, i.e. between 40 and 50. The Ward linkage method and Euclidean metric were used for this evaluation. The same performance measures were used for the second step, i.e. the Silhouette score, the Calinski-Harabasz score, and the Davies-Bouldin index.

The Silhouette score measures how compact and separate the clusters are thus, it was found to be between 0.16 and 0.18 for the chosen cluster range. The Calinski-Harabasz score shows how well-defined and separated the clusters are and this metric resulted as gradual decrease with an increase in the number of clusters. The Davies-Bouldin index, which is better when lower, assesses the effectiveness of clustering. It showed slight variations within the chosen cluster range.

The analysis yielded several key findings. Firstly, the Silhouette scores, stayed high (0.16 to 0.18) for cluster counts between 40 and 50. Notably, the highest scores (0.18) were at the range's extremes, both at 40 and 50 clusters. This indicates a strong degree of cluster cohesiveness and distinctness at these counts. Regarding the Calinski-Harabasz scores, a noticeable downward trend was observed as cluster numbers increased from 40 (998.68) to 50 (890.93). This suggests a slight but consistent reduction in cluster definition and separation with more clusters. Lastly, the Davies-Bouldin index, showed minimal fluctuation, ranging narrowly from 1.82 to 1.86. This stability indicates that the relative quality of clustering was maintained across the considered range of cluster quantities.

In summary, the detailed examination of clustering performance within the 40 to 50 cluster range, employing the Ward linkage method in conjunction with the Euclidean distance metric, has yielded comprehensive insights. The determination of the optimal number of clusters within this range presents a nuanced decision. A meticulous analysis of the results, considering the balance between the three metrics, points toward the selection of 40 clusters as the optimal choice.

Overall Data Structure based on CM

The dendrogram is generated from the hierarchical clustering of the TUS temp dataset as a byproduct of the grid search applied to agglomerative clustering. This provides a visual exposition of the data's stratification based on occupant demographic. The values on the x-axis of a dendrogram in hierarchical clustering typically represent individual data points if they are leaf nodes, or clusters of points if they are internal nodes. Figure 4.10 is a display of a high-level bifurcation, which is indicative of two major archetypal distinctions within the demographic data.

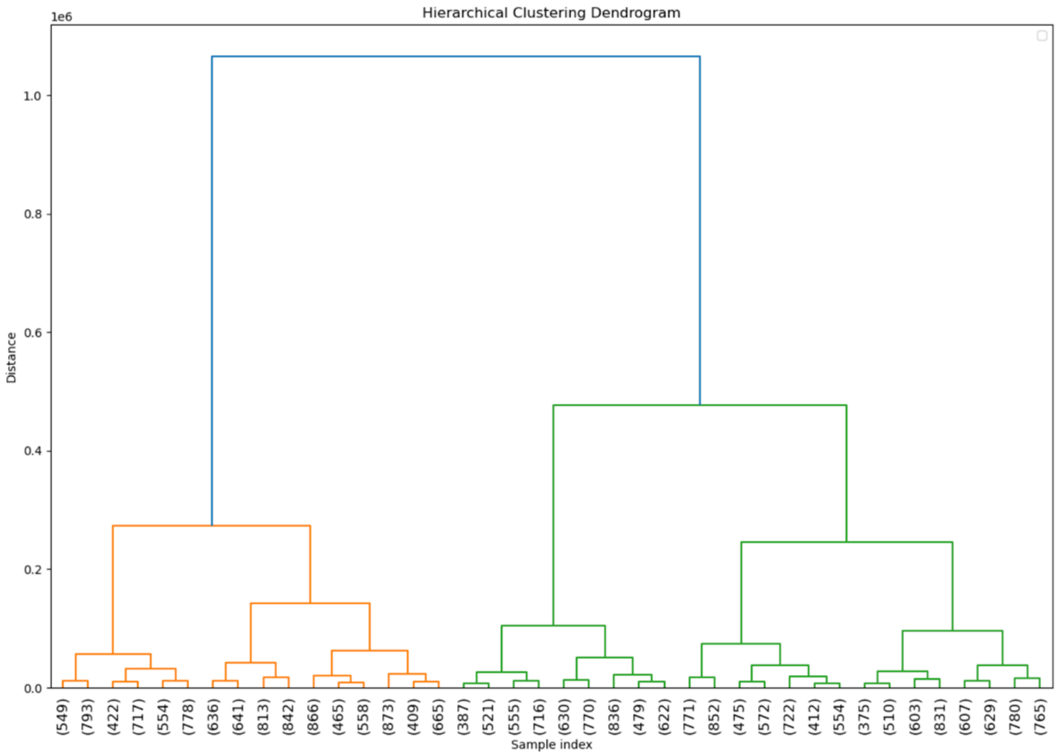

Figure 4.10. Dendrogram, the number of clusters by the levels of linkage

Subsequent divisions, particularly pronounced within the lower third of the distance scale, suggest the existence of more diverse archetypes. The clustering into 40 distinct groups reflects a detailed granularity in the profiling, with the length of the branches offering insight into the homogeneity of the archetypes. The dendrogram thus not only confirms the presence of varied demographic archetypes but also guides the selection of an appropriate number of clusters to represent them meaningfully, i.e. 40. Consequently, the archetype formation for the occupant demographics is supported according to dendrogram shape.

4.3.2 Feature Importance for Occupant Demographics

The feature importance analysis provides crucial insights into the variables that significantly influence cluster formation. Employment status emerges as the most influential factor, suggesting that the occupant's job status is a key determinant in differentiating occupancy patterns. "Age classes" and "number family members" also contribute which are indicating that both the age distribution within the household and family size play substantial roles in clustering.

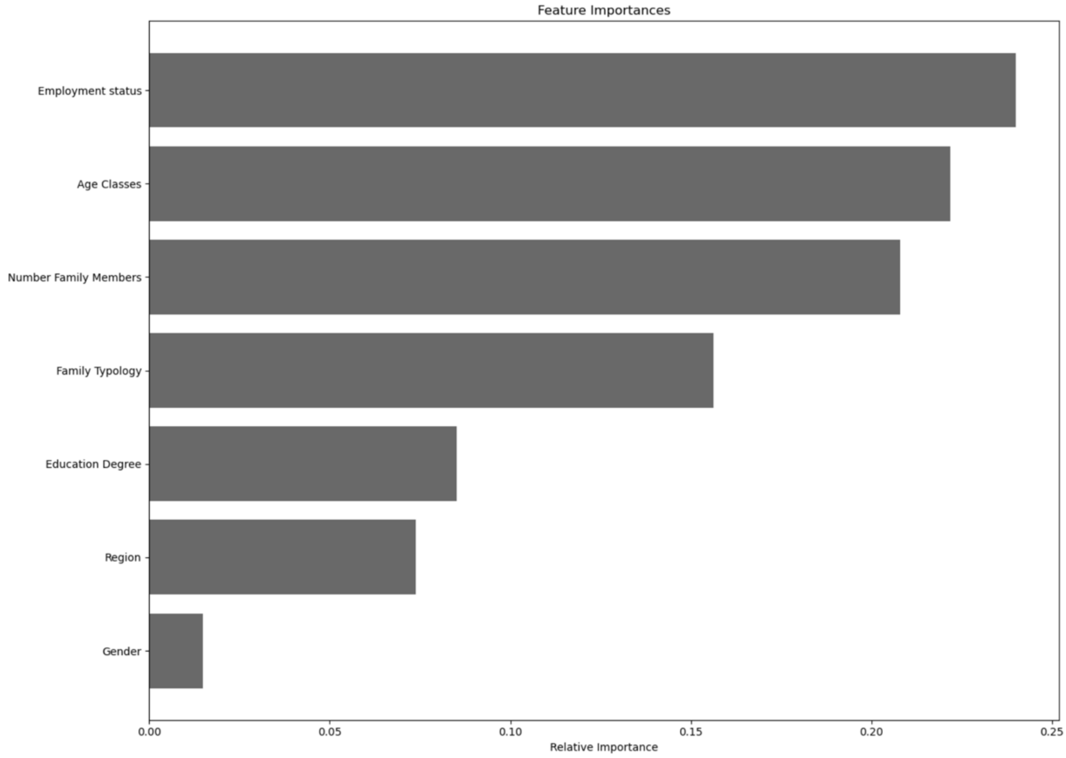

Figure 4.11. Feature Importance of the occupant demographic parameters

Notably, education degree and region hold moderate importance, reflecting their role in occupancy behaviors, while gender appears to have the least influence on the clustering outcomes as shown in Figure 4.11. This hierarchy of feature importance underscores the complex interplay of demographic characteristics in shaping residential energy demand patterns.

Evaluation of the Cluster Distinctiveness Using t-SNE Analysis for Feature Importance

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction technique, and it is mostly used in the field of ML for the visualization of high-dimensional datasets (Linderman & Steinerberger, 2019; Van Der Maaten & Hinton, 2008). The application of t-SNE in cluster analysis serves several roles: it enables the transformation of complex, multi-dimensional data into a two-dimensional space conducive to human interpretation, facilitates the assessment of cluster separation, and aids in the identification of inherent data patterns that are not discernible in the original high-dimensional space. These capabilities are particularly beneficial in evaluating the effectiveness of clustering algorithms for Hierarchical Clustering because they provide a visual confirmation of the data's organization into discrete clusters. Thus, the aim of the application of t-SNE is to test the clustering performance of the hierarchical clustering.

To find optimal t-SNE embeddings for the discernment of cluster structures, perplexity and number of iterations are the hyperparameters. Perplexity is a proxy for the number of effective nearest neighbors. It is varied across a range of [10, 30, 50]. The number of iterations is a parameter that dictates the extent of convergence. The range for this one was at [500, 1000, 2000]. These parameters collectively modulate the balance between capturing local and global data structures within the visual domain. Figure 4.12 represent both the tuning process and the best option which has the values of perplexity: 50, n_iter:2000.

The Calinski-Harabasz score peaked at lower cluster counts for 10 and 20. It decreased with larger cluster numbers, suggesting more defined clusters at these lower counts. Additionally, lower Davies-Bouldin scores were often seen with the Ward linkage and Euclidean metric. This indicated effective clustering. In conclusion, the analysis suggests an optimal range for the number of clusters (n_clusters) to be within 40-50. This range is chosen based on high Silhouette scores, showing well-defined cluster separation. There is also a balance in Calinski-Harabasz and Davies-Bouldin scores. This balance suggests optimal cluster density and minimal variation within clusters.

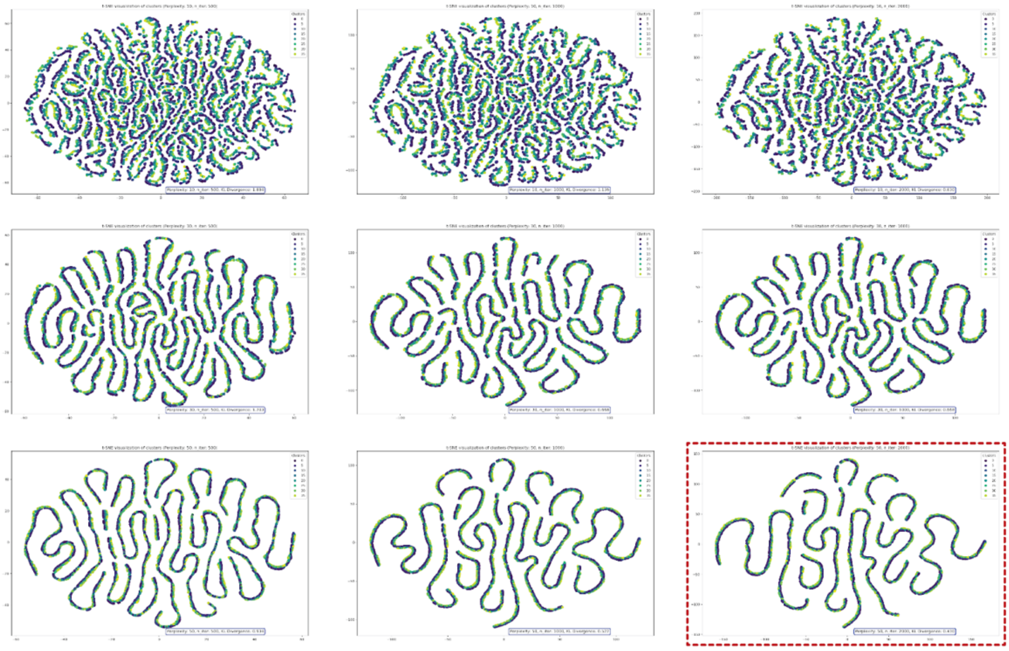

Figure 4.12. Tuning process of t-SNE, best option, right-bottom

In this analysis, the t-SNE algorithm has been parameterized with `n_components=2`, `perplexity=50`, `n_iter=2000`, and `random_state=42` to foster a balance between local and global data structure representation. The perplexity parameter is set in accordance with the number of clusters present. This can easily elucidate both micro and macro cluster relationships. The iterations are elevated to 2000 to ensure convergence to a stable solution. This is indicated with the value of 0.481of a Kullback-Leibler divergence, consequently, it presents a satisfactory embedding with minimal information loss.

The final t-SNE visualization visualizes a series of well-defined looping structures that denote cluster formations. These configurations suggest a high degree of intra-cluster similarity and inter-cluster dissimilarity. It can be interpreted as affirmation of the efficacy of the clustering process. The distinctive looping patterns observed may be emblematic of underlying cyclical trends within the dataset or could represent a mathematical artifact resulting from the t-SNE algorithm's optimization process. The visualization's consistency, despite variations in the `random_state`, implies a genuine feature of the dataset rather than a stochastic anomaly.

4.3.3 Imputation process of missing daily presence schedules with CM

After implementing t-SNE algorithms to validate the distinctiveness of clusters derived from hierarchical clustering, it was confirmed that the clusters are well-defined. However, it’s crucial to note that the augmentation applied to the activity and presence schedules with ML with feature clustering involved duplicating schedules rather than generating new ones. This approach operates by completing missing activity schedules from occupants within the same cluster, considering both seasonal variations and weekday categorizations.

The approach is not entirely accurate even previous studies has proposed this method for occupancy modeling. It represents a fundamental method of generating daily presence and activity schedules for year-round analysis but lacks comprehensiveness. In contrast, the end-to-end process, includes occupant interactions in the classification, providing a more realistic view. Conversely, the machine learning approach with feature engineering used here making it a basic representation and it only evaluates at the occupant level.

Example from the Individual Occupant Presence Schedules

The provided plot illustrates the synthesized presence schedules for a household (ID: 7740, randomly selected) across all four seasons and three distinct types of days: weekdays, Saturdays, and Sundays. The schedules are the culmination of a methodical imputation process employing Hierarchical Clustering. The schedules are based on the location column in the TUS dataset, i.e. “1” to denote the presence of at least one occupant and a "0" for their absence. This algorithm is applied to each hour of every day across the various seasonal and weekly categories.

Figure 4.13 is the graphical representation which underscores clear seasonal variations in occupancy. For example, during weekdays in winter, there is a notable presence in the early mornings and evenings, reflecting typical workday patterns. This contrast is observed with the extended presence at home during weekends, suggesting leisure or non-work-related activities. The occupancy during summer weekends shows a different pattern, potentially indicating variations in social activities or vacation behaviors during this season.

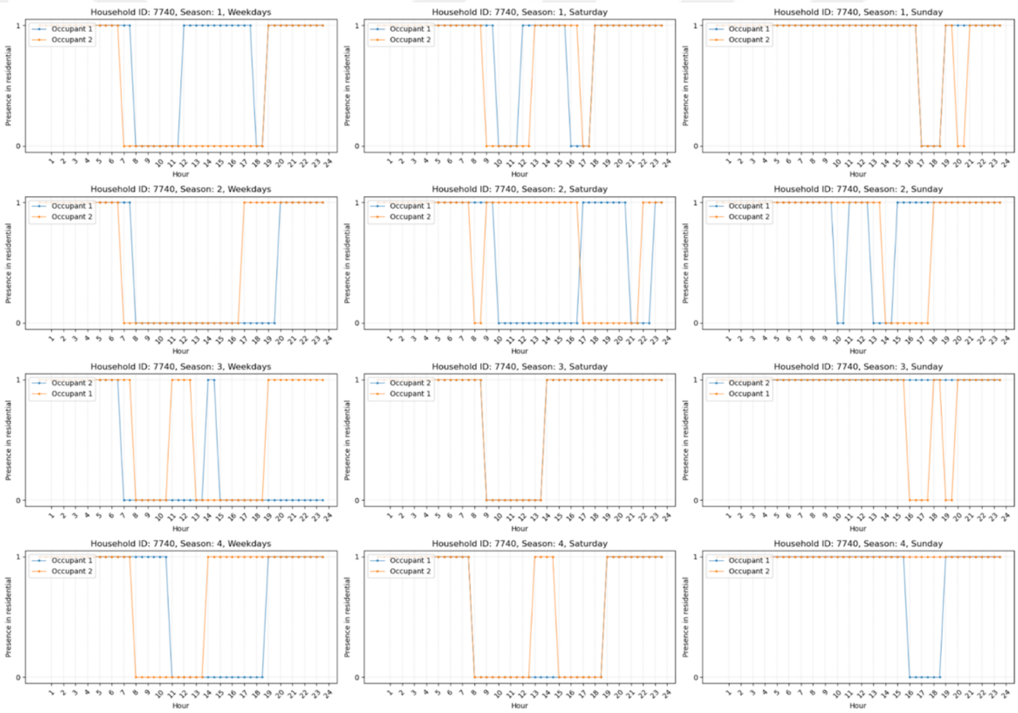

Figure 4.13. Presence schedule of a household: 4 seasons x 3 weekday category

The figure provided demonstrates the daily presence schedules for a household that have been augmented, showing variations across different seasons and weekdays. Despite the success in the clustering process, the post-augmentation analysis has indicated that there are significant discrepancies. This can be easily approved as the figure where the daily schedules seem reasonable in depicting sleeping and active hours. Nevertheless, it is apparent that there is a lack of consistency in the variation due to seasonal effects. For example, the expected increase in time spent outdoors during the summer is not clear, as the augmentation has resulted in what appears to be randomly generated outdoor hours. As has been explained in prior chapters, while the clustering method may be adequate for non-sequential demographic data, the generation of sequential presence or activity schedules necessitates the adoption of more advanced and holistic approaches, leading to the proposal of an LSTM model.

Example from the Individual Occupant Presence Schedules

Figure 4.14 depicts the imputed daily presence schedules for an entire year (i.e. four representative week schedules for each season and three weekday categories for week compositions). It suggests a discrepancy in the seasonal variation of an individual's presence schedules, despite implementation of demographic clustering.

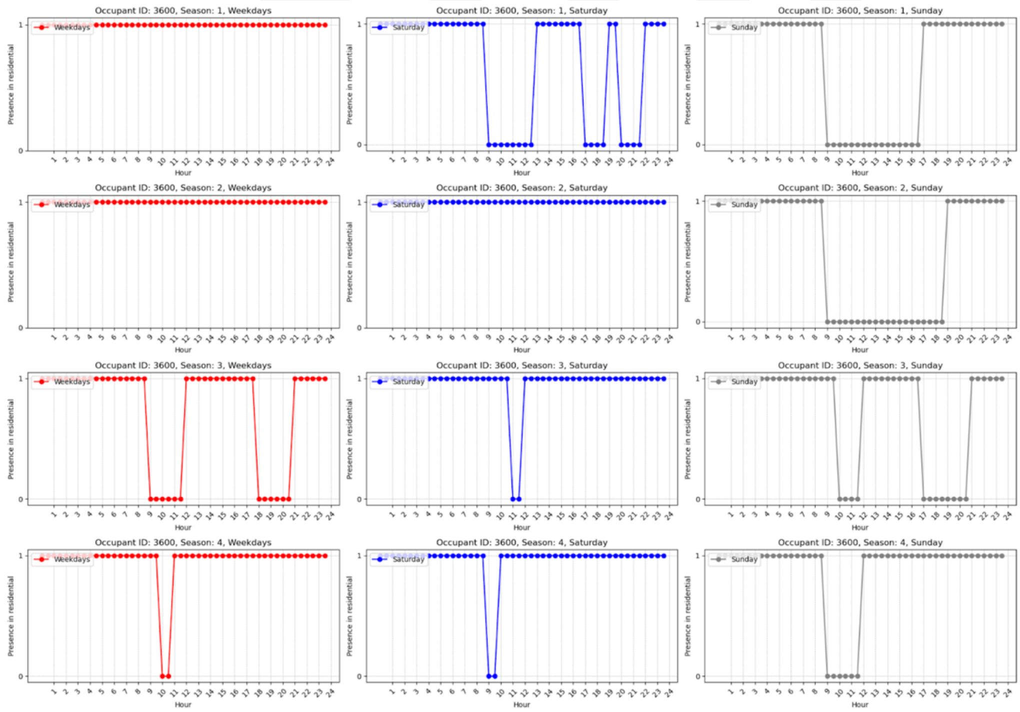

Figure 4.14. Presence schedule of an occupant: 4 seasons x 3 weekday category

The displayed patterns do not consistently reflect the lifestyle of a single person across different seasons, indicating a lack of uniformity in the modeled behavior.

The resulting schedules exhibit incongruences that suggest the need for a more nuanced approach in capturing the dynamic nature of individual occupant schedules throughout the year.

This imputation process does not consider crucial household information. This omission can lead to inconsistencies when these individual schedules are aggregated and applied to simulate entire households in building energy models. The potential confusion arises because the activities of occupants are interrelated and contextual within the household environment.

Clustering Results for Occupant's Presence Schedule: Statistical Differentiation Analysis

A similar chi-square test is implemented for the presence schedules that are generated with ML with Feature Engineering method.

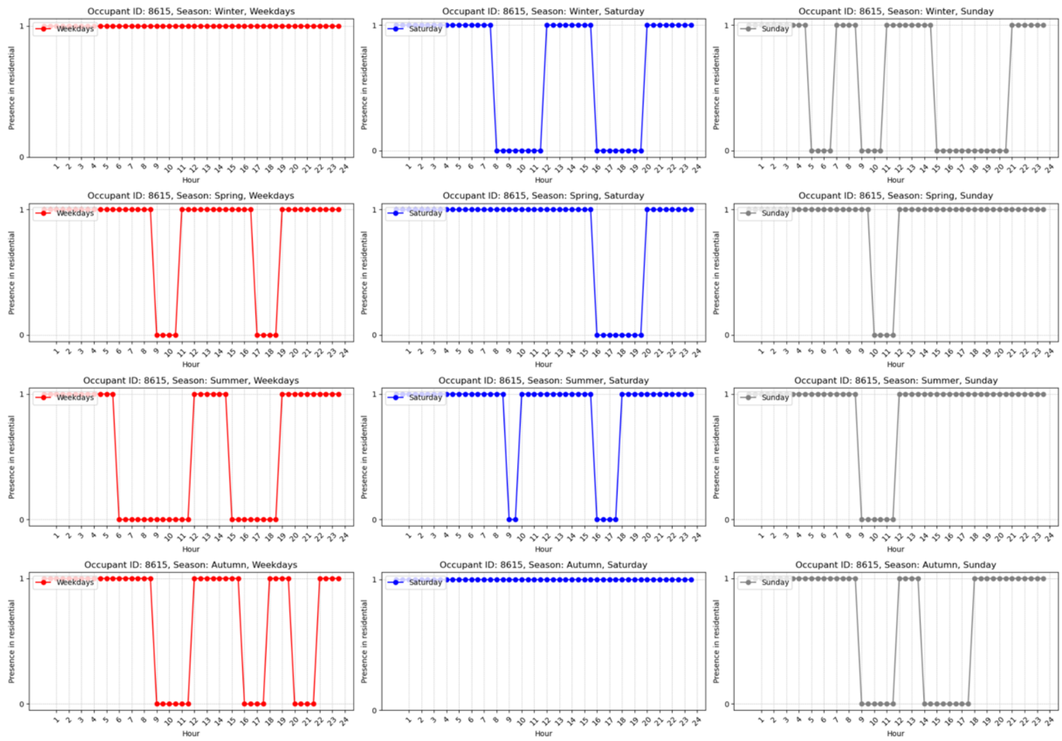

Figure 4.15. Chi-square test (i.e. p-value) of presence of an occupant

The methodology analyzes a binary presence schedule, where “1” indicates the presence and “0” signifies the absence of the occupant for whether the schedules are belonged to the same occupant across various seasons. The analytical results from the Chi-square test provide the measurement how the augmented schedules belong to same occupant as shown in Figure 4.15, i.e. 6.813 for weekdays, 0.00012 for Saturdays, and 0.00023 for Sundays. On weekdays, the relatively high p-value implies insufficient statistical evidence to reject the null hypothesis of a consistent distribution pattern, suggesting that the weekday presence schedules likely belong to the same individual. In contrast, low p-values for Saturday and Sunday indicate significant statistical deviations in the occupant’s presence on these days as compared to other times.

The discrepancy suggests a variation in the occupant’s routine or the involvement of a different individual during weekends. More examples can be found in Appendix C of the study.

Limitations of Clustering for Imputing Missing Schedules

This section highlights the disparities between the two methodologies and justifying the decision to adopt one over the other for the specific application of imputing missing daily schedules. It underscores the fundamental differences in model outputs and objectives about providing clarity on why the clustering-based ML approach falls short in the context of generating realistic, occupant-specific activity patterns.

The ML with Feature Engineering (i.e. clustering approach) leverages archetype profiling within demographic data to inform the filling of missing daily schedules. This approach is relatively straightforward to implement and provides interpretability in contrast to the “black-box” nature of LSTM models. to fill gaps for others, missing data can be quickly and coherently addressed. However, this method falls short of generating realistic daily schedules despite its simplicity and clarity. Therefore, it is important to consider individual differences in daily routines when generating schedules. It essentially replicates activity and presence patterns from one occupant to another within the same cluster by assuming homogeneity in daily routines among similar demographic groups. While this might capture a broad-strokes representation of occupant behavior, it does not account for the individual hourly variations and unique daily activities of each occupant. The clustering approach may overlook differences in daily routines that occur not only between different occupnats but also for the same occupant across various days and seasons.

The clustering approach operates on the level of occupant demographics, which inadvertently overlooks the broader context of the household. In contrast, LSTM models consider the household as an integral component, capturing the interplay of occupant activities within this shared living space. This omission is significant because occupants' daily routines are often interlinked, with household activities being collaborative or influenced by other members. Thus, the augmentation process should consider the rich tapestry of inter-occupant relationships and their impact on daily routines.

While the clustering-based ML approach does incorporate some level of household context through the “Family Typology” column, its effectiveness in capturing the full spectrum of household dynamics remains limited. This column categorizes families into types such as “Lonely person,” “Childless couples,” “Couples with children,” and other family structures, providing a basic framework of the household composition. However, this categorization primarily focuses on demographic aspects and does not delve into the daily interactions and activities of the household members. Consequently, even with this added dimension of family typology, the clustering method predominantly centers on individual demographics, leading to a shortfall in accurately modeling the nuanced and dynamic interactions that influence daily routines in a household setting.

Lastly, as demonstrated in Figure 4.13 and Figure 4.14, it is revealed that accuracy in capturing seasonal or weekly variations is compromised when augmenting the TUS dataset via clustering-based ML approaches. This process is flawed because its evaluation isn't thorough enough. The only thing being evaluated is the number of clusters not specifically individual level for each occupant. In contrast, within the LSTM approach, individual variation for each occupant is facilitated by the deep learning algorithm's capacity to integrate both sequential and non-sequential data. Such a level of resolution is deemed necessary for precise occupancy modeling at the occupant level.

Therefore, ML with feature engineering cannot be relied upon for schedule augmentation, as it only redistributes existing schedules rather than creating new. Consequently, these limitations highlights the need for a more comprehensive approach that not only acknowledges the family composition but also integrates the complex patterns of interaction and activity within the household, essential for a realistic and individualized understanding of occupant behavior.

4.4 Experiment 2: Stochastic Method (SM): Markov-Chain Sequences

This section conducts a detailed comparison between the proposed model and other prevalent models, with a focus on the Markov-chain approach. Each model offers a distinct perspective on occupancy generation for presence schedules with varying levels of predictive capability. The LSTM model is with its deep learning architecture and embedded layers that integrate week, season, and demographic features, is set against traditional MC models. These demographic features are also implemented in the markov models using contextual transition matrices. However, as it is explained in the pervious chapters, these models can not sucessfully handle the multi-categorical classification tasks. In parallel, these conventional models have been widely utilized in existing research for generating occupancy schedules based on presence schedules. Thus, the comparison between two models is executed in terms of presecnce schedule predictive capacitiy as these schedules consist of only binary states, ie., present or absent in the residential.

Furthermore, the Markov models usually operate under the assumption of the Markov property. The prediction of the next state depends solely on the current state, not on the sequence of events that preceded it. This contrasts with the LSTM model, because it can processes the entire sequence. Therefore, it can capture the long-term temporal dependencies or extremities within the data. Consequently, this comparison is intended to showcase the limited capabilities of Markov-chain models for comprehensive and realistic occupancy modeling.

Single Example: First order & No Contextual & Active Initial State

The generation of occupancy schedules is exemplified through the application of a first-order Markov chain. The presented model is incorporating an "initial_state" and depicted through a 48-row visual representation. The model performance is evaluated the model's accuracy based on three metrics: hamming distance, jaccard similarity, and accuracy. This model's predictive efficacy can be interpreted by various factors, i.e. occupant's employment status, the day of the week, and the frequency of state transitions. The model performs well during weekdays for employed individuals or students because they exhibit regular patterns of presence. In contrast, schedule of non-working occupants may vary as they do not have to regularly leave the house. The regularity of tbeing employed or student decreases on weekends. Because the routine behavior of all residents becomes less predictable. The model's predictive accuracy does not change significantly with an increase in the number of transitions which can be multiple departures from residential at one single day.

The Figure 4.16 presents a daily presence schedule which is structured into 30-min. This reveals the occupant’s location status as either at home ('1') or away ('0'). The "HHID 5267" is an actual schedule which is used as a template for first-order Markov-chain process. The schedule is representing an occupant who is employer and who is leaving the residential regularly during weekdays.

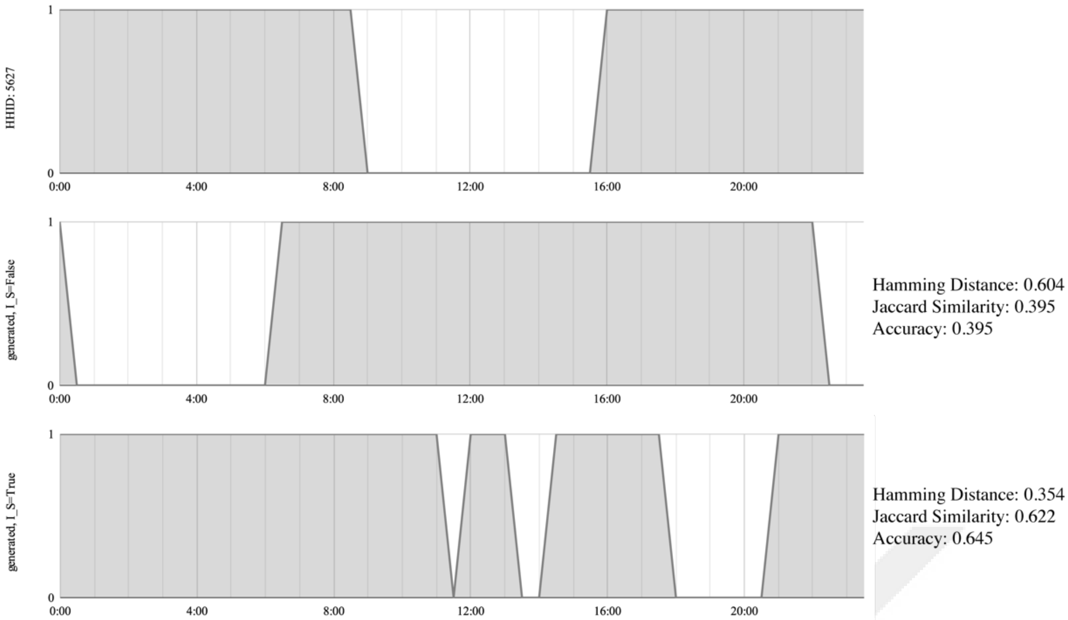

Figure 4.16. First-order markov-chain single generated presence schedule example, single state transition

The "Generated 1 - False" segment demonstrates a generated schedule that is based on first-order Markov-chain process without initial state information. Initial state is the feeding the Markov process giving the first element(s) of the schedule list. In this example it is automatically selected from the actual presence schedule of HH_ID 5627. The "Generated 1 - True" schedule displays a schedule that aligns closely with the real data, and this indicates the model performance is significantly enhanced when the model is furnished with the initial state. This improvement emphasizes the importance of incorporating initial state data in the model to better capture the probabilities of transition between states, thereby augmenting the model's overall accuracy in reflecting true occupant behavior.

These graph indicate that incorporating the initial state into the Markov chain model can enhance prediction performance, even when facing schedules with multiple transitions. Contrary to the assumption that an increase in the number of state changes invariably leads to reduced predictive accuracy, this case demonstrates that increase in number of transitions does not necessarily diminish the model's performance. Finally, the initial state serves as a critical anchor for the predictive process enabling the Markov chain to adapt to the occupant's daily presence rhythm.

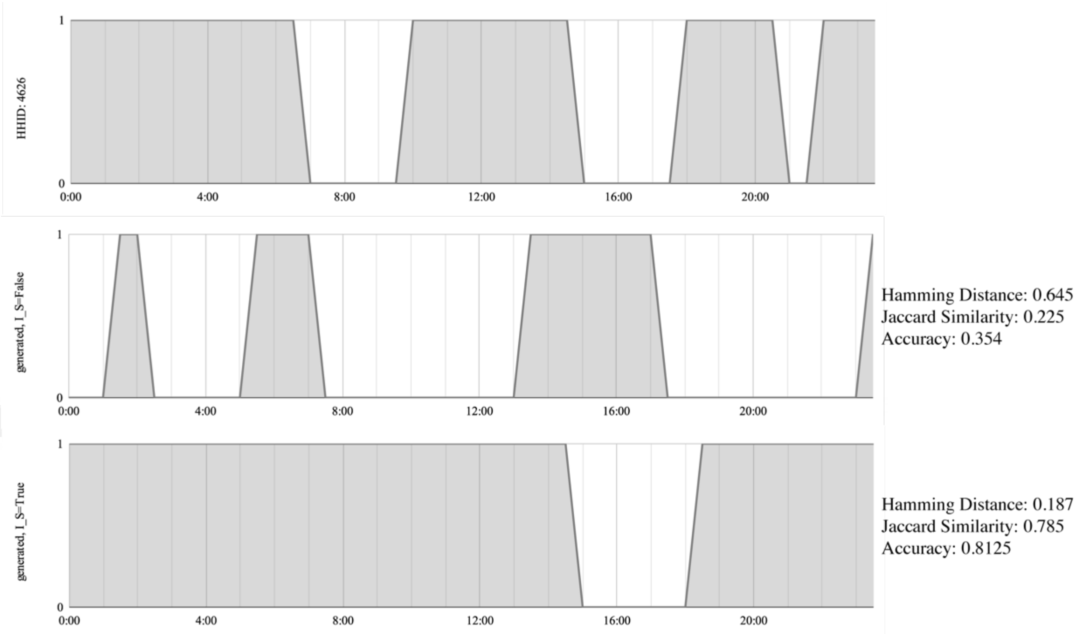

Figure 4.17. First-order markov-chain multiple generated presence schedule example, multiple state transition

Figure 4.17 exhibits a presence schedule with multiple state transitions within a day. The first part is actual schedule of HH_ID 4206. The schedule consists of several instances of the occupant leaving and returning home, indicative of a dynamic and non-routine lifestyle. The "Generated 1 - False" depicts an incorrectly predicted schedule by the Markov chain model as it is not aligning well with the actual transitions. However, the “Generated 1 - True" suggests a successful prediction of the occupancy state despite the increased number of transitions. This can be counterintuitive to the common expectation that more transitions would complicate the prediction.

Probabilistic Nature of Markov Chain and Need for Providing Sufficient Data for Testing

The observation on single trials of a MC model can yield varied outcomes due to model's intrinsic probabilistic nature. Each execution of the model yields distinct outcomes which are governed by probability distributions rather than deterministic rules. This probability distributions are generated based on the transition matrices which is formed specifically for the dataset as shown in Figure 4.18. This characteristic is particularly pronounced for higher order because the extended state sequences amplify the variability due to less frequent occurrences of specific sequences in the data. Therefore, it is crucial to provide a robust dataset in terms of substantial both in volume and diversity.

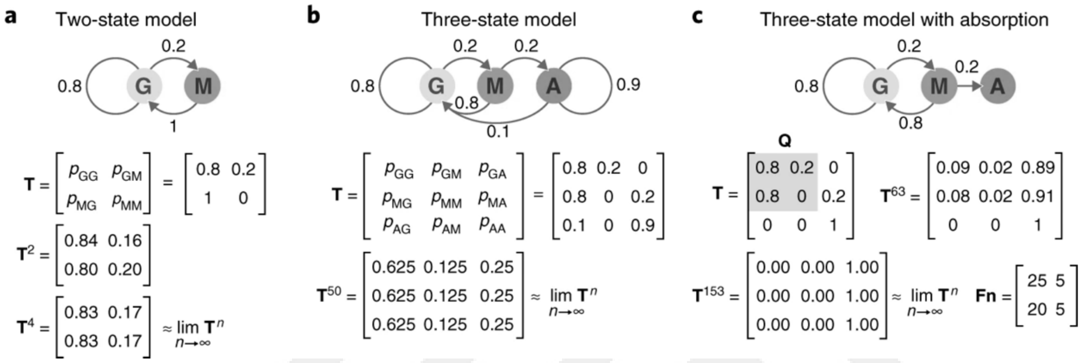

Figure 4.18. State transition models based on the numebr of the states, transition matrices T (Grewal et al., 2019)

The adequacy of the data is very important; It supports the reliability of the model by ensuring that the derived probabilities are representative of the underlying patterns, thus reducing variation in repeated predictions and increasing the model's prediction accuracy. In detail, here are the key reasons behind this variability include:

- Randomness in State Transitions: The transition from one state (or state sequence) to the next is probabilistic thus, it is leading to potential variations in the generated schedule.
- Initial State Selection: Use of “initial state” for the generated schedule adds another layer of randomness.
- Model Characteristics: Higher-order Markov chains introduce more variability due to consideration for multiple previous states.

Comparison based on Markov-Chain Order for all the Dataset

The Markov chain process is applied across the entire dataset as in the previous chapter is stated single examples are not enough to evaluate Markov generations. The application consists of first, second and high order (i.e. third order) with as the variations in the order and the incorporation of an initial state for each order alternative. The influence of these variations is examined in terms of the predictive performance of presence schedules.

Table 4.3 Comparison in terms of order of markov-chain with intial state option

	First order	*First order with IS**	*Second order*	*Second order with IS*	*High order*	*High order with IS*
Hamming Distance	0.450	0.383	0.4481	0.399	0.461	0.380
Jaccard Similarity	0.489	0.564	0.483	0.545	0.471	0.560
Accuracy	0.549	0.616	0.551	0.600	0.538	0.619

*first-order when initial state is true during generation

Table 4.3 presents a comparison of Markov chain models of different orders with and without the use of an initial state. The evaluation metrics used for comparison are Hamming Distance, Jaccard Similarity, and Accuracy. Three metric is selected due to more robust performance evaluation.

<u>First Order, Initial State False:</u> This baseline model shows the highest Hamming Distance (0.450), and the lowest Jaccard Similarity (0.489) and Accuracy (0.549). These results point out that this setup is less effective for predicting occupant presence, i.e. 48-rows of binary presence states.

<u>First Order, Initial State True:</u> By incorporating the initial state, there is a Notable improvement is observed with the implementation of initial state which is the feeding the Markov model with the first element of the actual schedule. Hamming Distance decreases to 0.383, Jaccard Similarity increases to 0.564, and Accuracy to 0.616.

This demonstrates that knowing the initial state significantly enhances model performance.

Second Order, Initial State False: The predictive performance is slightly better than the first order without initial state as Hamming Distance is 0.4481, Jaccard Similarity as 0.483, and Accuracy as 0.551. This suggests that simply increasing the order of the Markov model without considering the initial state do not offer marginal improvement for the presence schedule performance.

Second Order, Initial State True: There's an improvement across all metrics compared to its counterpart without the initial state: Hamming Distance drops to 0.399, Jaccard Similarity climbs to 0.545, and Accuracy to 0.600. The initial state continues to show its impact for different order levels.

High Order, Initial State False: This model performs worse than the first-order model with no initial state, with the Hamming Distance increasing to 0.461, and both Jaccard Similarity (0.471) and Accuracy (0.538) decreasing. This result suggests that higher-order complexity without the initial state may not be beneficial. However, there could be another explanation. The previous part of the text mentioned that the model takes chance (probability) into account. So, the small changes in accuracy (less than 5%) might not be very significant.

High Order, Initial State True: This model achieves the best metrics: Hamming Distance is at its lowest (0.380), while Jaccard Similarity (0.560) and Accuracy (0.619) are the highest. As a result, it is indicated that incorporating the initial state at a higher order yields the best predictive performance.

The data suggests that while increasing the order of the Markov chain model could not offer improvement in predictive performance. On the other hand, the inclusion of the initial state is a more significant factor in enhancing the model. Higher-order models without the initial state do not necessarily provide better predictions, the immediate predecessor's state might be the most influential factor. Consequently, the implementation of the initial state across all orders consistently improves the model's accuracy. Lastly, during the modeling process, in cases where no transitions are observed in the generated schedules, a fallback strategy is employed. This involves equally distributing the probability across all possible states that is commonly referred to as smoothing for managing occurrences of events that haven't been seen before in probabilistic models.

Comparison Based on Implementation of Occupant Demographics

A comparative analysis with the integration of contextual parameters into higher-order (i.e. third order) Markov-chain models is conducted for occupancy presence schedules. As the in the previous comparison, higher degree Markov chain performed the highest accuracy thus, the setup is formed based on this model. Similar with the previous chapter, the comparison is based on predictive performance of the Markov model for three metrics. Three alternatives of the higher-order Markov-chain model is applied, i.e. the first without any contextual parameters, the second incorporating periodical features from the TUS daily dataset, and the third encompassing both periodical features and extensive occupant demographics. Periodical features are “months_season” and “week_or_weekend” parameters and occupant demographics are "Region", 'Number_Family_Members', 'Family_Typology','Age_Classes', 'Employment_status', 'Education Degree'.

Table 4.4 compares different occupancy models based on their contextual setup and the implementation of the initial state (IS). The evaluation metrics used for comparison are Hamming Distance, Jaccard Similarity, and Accuracy. Three metric is selected due to more robust performance evaluation. *No-Contextual with Initial State False* and No-*Contextual with Initial State True* setups are the based models of this comparison and they have explained in the previous chapter as the names of *High Order, Initial State False* and *High Order, Initial State True*.

Table 4.4 Comparison in terms of implementing the contextual parameters to the markov-chain model with option of intial state

Higher-order (i.e. third)	*No-Contextual*	*No-Contextual with *IS*	*Partial Contextual with IS*	*Partial Contextual with IS*	*Full Contextual with IS*	*Full Contextual with IS*
Hamming Distance	0.461	0.380	0.428	0.344	0.374	0.308
Jaccard Similarity	0.471	0.560	0.520	0.592	0.516	0.600
Accuracy	0.538	0.619	0.571	0.655	0.625	0.691

*first-order when initial state is true during generation

Partial Contextual, Initial State False: This model integrates partial contextual information along with the initial state. It shows better metrics than the no-contextual models, with a Hamming Distance of 0.428, Jaccard Similarity of 0.520, and Accuracy of 0.571. This indicates the added value of including some context in the modeling process.

Partial Contextual, Initial State True: Although the same category as the previous, it appears to show improved metrics with a Hamming Distance of 0.344, Jaccard Similarity of 0.592, and Accuracy of 0.655. This improvement suggests that even within 'partial contextual' models, there can be variations that affect performance.

Full Contextual, Initial State False: This model shows further improvements as the metrics are a Hamming Distance of 0.374, Jaccard Similarity of 0.516, and Accuracy of 0.625. It seems there's a discrepancy in the expected trend with Hamming Distance increasing slightly compared to the partial contextual model.

Full Contextual, Initial State True: Representing the most advanced model, it has the best overall metrics: Hamming Distance is at its lowest (0.308), Jaccard Similarity at its highest (0.600), and Accuracy peaks at 0.691. This demonstrates the powerful impact of full contextual information combined with the initial state on the model's predictive ability.

As the models progress from no contextual through partial to full contextual integration, a trend of performance enhancement is observed. Each incremental step toward including more context, especially when it is combined with the initial state, there is a consistent reduction in Hamming Distance. This is accompanied by an elevation in Jaccard Similarity and an increase in Accuracy. These improvements underscore the substantial benefits of integrating contextual data into occupancy modeling. Nevertheless, it is noteworthy that even with the integration of full contextual parameters, the higher-order Markov chain models do not achieve the accepted confidence interval of 95%. This fact underscores the success and superior reliability of the proposed LSTM model.

4.5 Experiment 3: End-to-End Deep Learning Method (EM): Multi-task

The EM is the employment a Long Short-Term Memory (LSTM) network for specifically designed to classify daily activity schedules and the location of activities (i.e. presence schedule) for occupants in urban residential buildings. The model processes data at 30-minute intervals which is resulting in 48-time intervals per day. The model executes multi-categorical classifications for each period of 30-minute. The primary objective of utilizing the LSTM model is to expand the limited scope of the original CENSUS dataset, which contains only a singular daily activity schedule for each occupant. Then, it is aimed to generate twelve distinct schedules per occupant by accounting for three types of days within a week across four seasons. Consequently, it is aimed to produce a diverse and comprehensive set of activity schedules that accurately reflect the occupants' routines throughout the year and to classify the location of activities for each time interval, thus providing a realistic representation of daily life across various seasons and weekdays. As one of the original features, the model is designed to classify all 48-time intervals concurrently, focusing on capturing the sequence of daily activities, rather than merely predicting the final timestep.

Input Dataset for EM: TUS temporal

The augmentation of the dataset from a singular activity schedule per occupant to a more comprehensive assembly of 12 schedules. This expansion is designed to encapsulate a diverse array of daily routines by incorporating variations across four seasons (e.g. 1: Winter, 2: Spring, 3: Summer, 4: Autumn) and three types of days (i.e. diary days) within a week—weekdays, Saturdays, and Sundays.

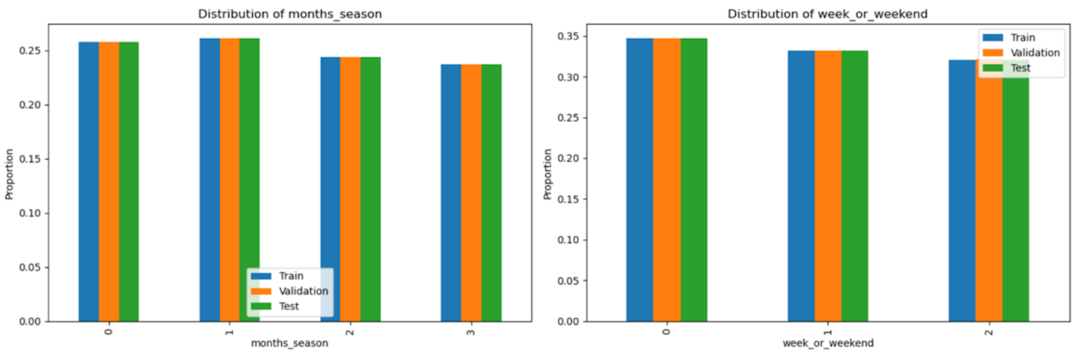

Figure 4.19. Activity schedule distribution based on seasonal/weekly variations

Figure 4.19 displays a detailed breakdown of the distribution of daily activities of occupants which is delineated by both seasonal and weekly cycles. This breakdown is pivotal for performance of LSTM model because as it can bee seen, the distributions between seasons and weekday categories are approximately equal.

Implementation: LSTM Model Training Phase

The training dynamics of the LSTM model augmented with embedding layers are shown in the Figure 4.20. Both the training and validation accuracies show a convergence trend, with the validation accuracy closely following the training accuracy. This suggests a balanced learning process without significant overfitting to the training data. At epoch 249, the model achieves a training accuracy of 68.36% and a validation accuracy of 68.49%, indicating a modest learning performance. In this setup only the periodical features, i.e. "months_season", "week_or_weekend" were included in the process of training as embedding layer parameters. Before tuning process, the occupant demographics are added besides periodical features to support the learning process, i.e. “Region”, “Number Family Members”, “Family Typology”, “Age Classes”, “Employment status”, “Gender”, “Education Degree”.

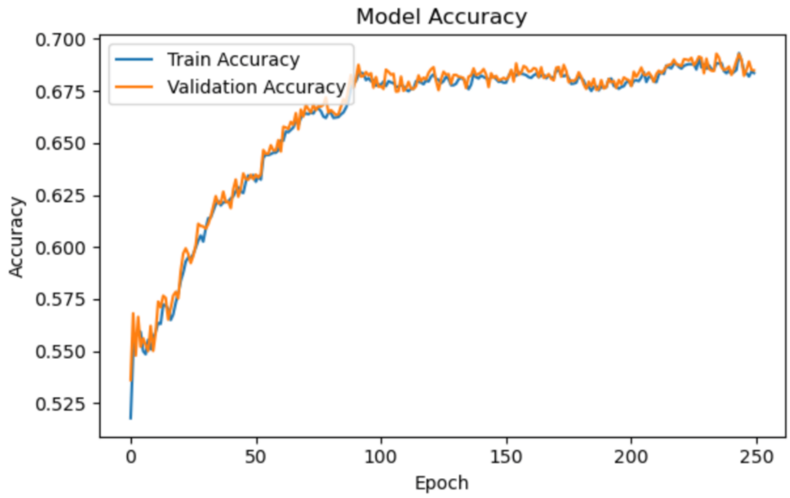

Figure 4.20. Training curve of LSTM, only periodical features are included in the training

Despite the progressive learning trajectory, there is a discernible margin for improvement to reach the aspirational benchmark of 95% accuracy for both the training and validation phases. This gap would require a strategy that includes hyperparameter tuning, model architecture optimization, and strengthening of generalization capabilities. This process is explained in detail in the results chapter.

Implementation and Training Details

The implementation is employed in HT for model training. This method usually is selected as well-suited for minimizing functions, particularly black-box functions with high evaluation costs.

The process of HT involves several key steps. Initially, the objective function is defined which trains the model using specified hyperparameters and returns the desired metric, e.g. validation loss, accuracy. Subsequently, the hyperparameter space is established for delineating the range or set of values for each parameter under consideration. Then, a surrogate model is then selected from the alternatives, e.g. Gaussian Processes, Random Forests, Tree-structured Parzen Estimators.

The optimization algorithm of TPE determines the next hyperparameters to be evaluated. It searches a balance between exploration of new hyperparameters and exploitation of the most promising ones. The acquisition function scores the hyperparameters proposed by the surrogate model as employs strategies i.e., like Expected Improvement (EI), Probability of Improvement (PI), or Upper Confidence Bound (UCB). Finally, the optimization is then executed.

Implementation: BO for LSTM with Embedding Layers

The study aimed to identify an optimal set of hyperparameters that would enhance the model's predictive accuracy, thereby ensuring robust and reliable performance in its application domain. The *Optuna* was configured to run for a predetermined number of trials, with each trial representing a complete training cycle of the model using a unique set of hyperparameters. Here are the key parameters and their ranges for tuning process.

- Number of hidden layers: [5, 6, 7, 8, 9, 10]
- Number of neurons in the hidden layers: [50, 100, 150, 200]
- Batch size: [48, 96]
- Learning rate: [0.001,0.0005,0.0001]
- embedding size: [50, 100]
- optimizer: ["adam", "rmsprop", "sgd"]
- initializer: ["glorot_uniform", "he_normal", "lecun_normal"]
- activation function: ["tanh", "relu"]
- selection of bidirectional for LSTM: [True, False]

Differentiation Analysis for Results of Presence Schedule by EM

An independence test, ie. Chi-Square, is developed as a method for assessing the consistency of an individual’s presence within a residential unit across different seasons. This approach involves organizing the data into a contingency table, with one axis representing different seasons and the other indicating presence or absence.

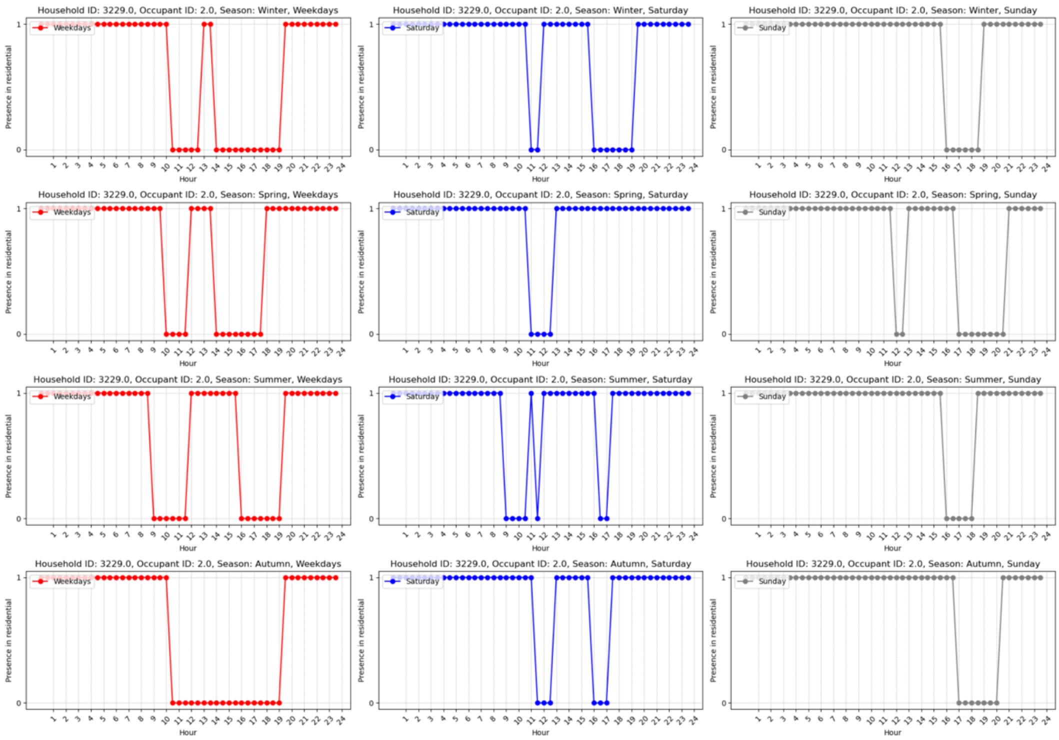

Figure 4.21. Chi-square results (i.e. p-value) of presence schedule of *occupant x*

The test evaluates the p-value; a value below the conventional threshold of 0.05 would suggest significant differences in presence patterns, thus, this could be interpreted as the schedules may not belong to the same occupant. The objective is to apply a statistical test to analyze whether these 48-point daily presence schedules (ie. binary) are characteristic of the same individual across various seasons.

The Chi-square test results for the occupant’s presence schedule reveal p-values of 0.5267 for weekdays, 0.5099 for Saturdays, and 0.5046 for Sundays (Figure 4.21). These values suggest a high degree of consistency in the presence schedule of the occupant across the entire week. The relative uniformity of the p-values indicates that there is no substantial statistical difference in the distribution of the occupant's presence between weekdays and weekends. This implies that the patterns observed throughout the week can reasonably be attributed to the same occupant.

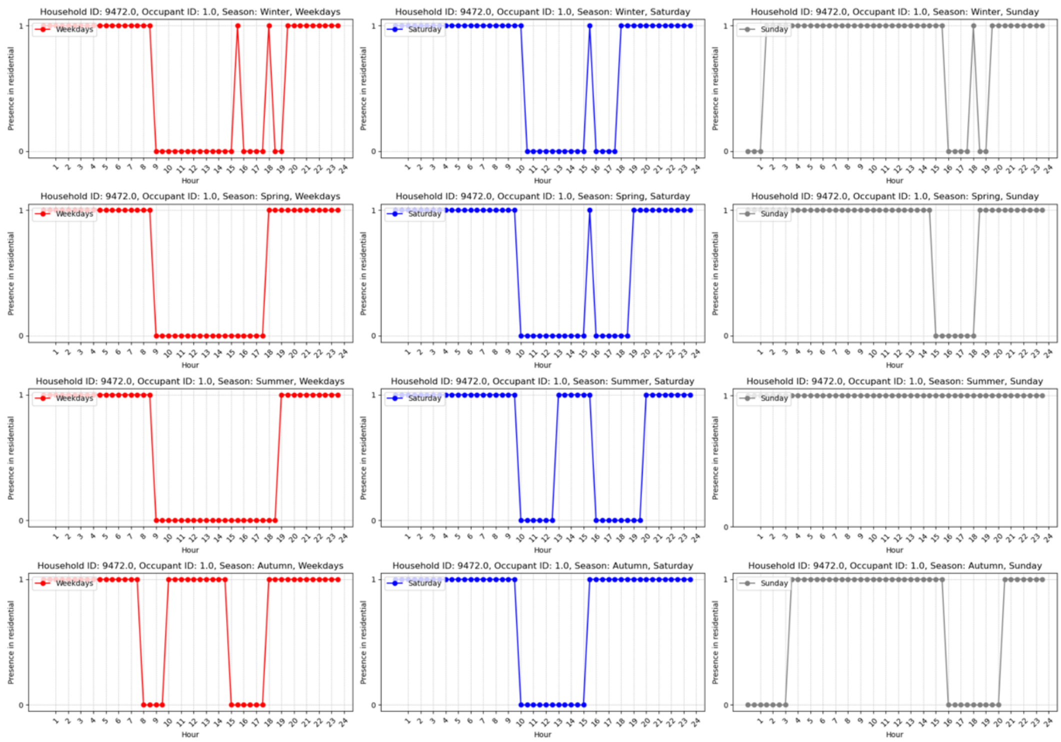

Figure 4.22. Chi-square (i.e. p-value) of presence schedule of an *occupant y*

Figure 4.22 represent insightful implications about evaluating the effectiveness of the proposed LSTM model for generating presence schedules. This can be a great example for how human behavior can be explained with patterns. Thus, p-values for the occupant's presence schedule are 0.12419 for weekdays, 0.61164 for Saturdays, and a notably low 0.00021 for Sundays. The results for weekdays and Saturdays indicate a successful augmentation. This success can be attributed to the presence of more consistent patterns and repetitions on these days, i.e. binary presence reflection of the occupant's routine and demographics. In contrast, the low p-value for Sundays underscores the model's limitations in accurately capturing presence schedules for this day. The unique nature of Sundays is a holiday, and this day results can entail a greater variation in terms of occupant and household behavior. This variability makes it challenging for the LSTM model to generate an accurate and consistent presence schedule for Sundays. This test shows why it's important to account for different behaviors on different days when using machine learning to predict if people are present.

Consequently, End to End Deep Learning Model without Feature Engineering methods succesfullt can reflect a stable and predictable routine in their occupancy behavior. The success in generating a reliable presence schedule and applying the Chi-square test underpins the effectiveness of this approach in analyzing occupant presence patterns in residential settings.

Classification Performance Beyond Accuracy for Results of Activity Schedule by EM

This chapter represent the multi-categorical classification performance of proposed LSTM model with embedding structure in terms of precision, recall, and F-1 score. These metrics is applied to 146 distinct activity types for "Occupant_Activity" column as addition to conventional accuracy score during the training process. This approach is adopted to provide a more comprehensive evaluation of the model's performance. Accuracy is offering an overview of overall correctness; however, it is insufficient for providing in-depth insights into class-specific performances or for addressing the challenges posed by imbalanced datasets and complex categorizations. Thus, the effectiveness of the model in correctly predicting positive instances is assessed using precision and recall. These metrics are important when the implications of false positives and false negatives exist. The F-1 score is calculated as the harmonic mean of precision and recall, and it offers a balanced metric for assessing performance across classes of diverse sizes and importance. The number of actual occurrences of each class, denoted as support, is also measured to provide an understanding of the accuracy of the model in relation to the distribution of the dataset. Subsequent part of this sub-chapter presents these metrics graphically.

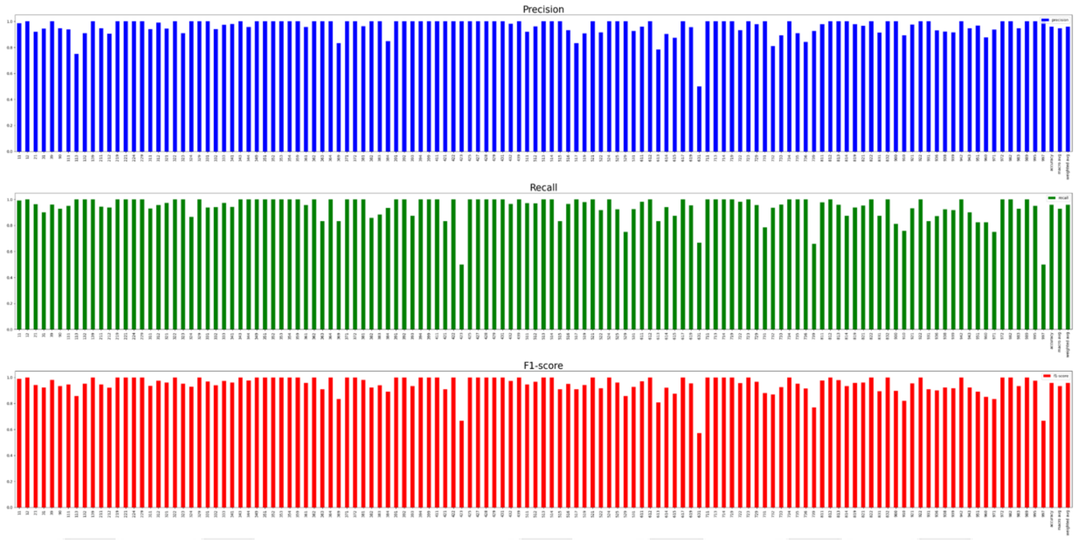

Figure 4.23. Performance metrics for all categories of “Occupant_Activity”

Figure 4.23 illustrates the classification performance of LSTM network across multi-categorical activities in terms of precision, recall, and F1-scores. The bar graphs display the values for each class the target variable.

- The precision graph (in blue) reveals a high degree of accuracy in the model’s positive predictions for indicating that when it predicts an activity. The results are high reliable for most of the activity category.
- The recall graph (in green) indicates the model’s strength in capturing all relevant instances of each activity. The generally high bars suggest that the model has a robust ability to detect the presence of the various activities without overlooking many actual cases.
- The F1-score graph (in red) reflects a harmonic balance between precision and recall; thus, it is pointing out that the model maintains a strong performance even when accounting for the trade-off between these two metrics.

For overall metrics, the model’s overall accuracy is at 95.9% which represents robust performance. The macro average is a simple average of the metric across classes, and it displays impressive scores as well as with precision at 94.8% and recall at

92.7%, and F1-score at 93.4%. These figures suggest that the model is not only accurate overall but also maintains this high level of performance consistently across the various classes. The weighted average is for class imbalance by weighting score of each class for its presence in the dataset. In summary, the LSTM classifier exhibits a strong and consistent performance across all metrics, signifying a successful multi-categorical classification effort. In detail representation can be found in Appendix E.

4.6 Experiment 3a: EM vs Various Embedding Structures of EM

The model presented is characterized by the integration of occupant demographics with embedding structures, and also incorporating periodical features into this embedding methodology. The amalgamation of temporal sequence datasets with non-sequential occupant demographics and non-sequential periodical features is central to the comparison outlined in this chapter. This blending of non-sequential and sequential elements represents one of the innovative aspects of the proposed LSTM model. Consequently, a comparison is conducted utilizing the LSTM framework to showcase the impact of the comprehensive occupant demographics and periodical features.

Three LSTM models, each with a distinct embedding structure, are analyzed comparatively:

- the first model (i.e. without embedding structure) is based on only continuous features e.g. cyclical encoded of hourStart_Activity, hourEnd_Activity.
- the second model (i.e. embedding with partial categorical columns) merges continuous features with periodical non-temporal categorical variables using embedding structure technique, i.e. "months_season" and "week_or_weekend". These columns are coming from the TUS daily sub-dataset thus they are isolated from the occupant demographics in the comparison.

- The third model (i.e. embedding with full categorical columns), which is the primary focus of the augmentation process of daily activity schedules, incorporates a comprehensive array of non-temporal features, including both periodic elements and detailed occupant demographics.

The Comparison Setup of Different Embedding Structures

The comparison of the three LSTM models was rigorously conducted based on their training performance using the best parameters identified through Bayesian optimization. The tuning process is executed to ensure that the selection of the best parameter values for the best performance for each comparison element. These parameters included as follows:

- the number of hidden layers (`num_hidden_layers`)
- units in each layer (`hidden_units`)
- batch size (`batch_size`)
- learning rate (`learning_rate`)
- embedding size (`embedding_size`)
- optimizer (`optimizer`)
- initializer (`initializer`)
- activation function (`activation_function`)
- the choice of bidirectional layers (`bidirectional`).

The training validation process evaluated both accuracy and loss metrics focusing on enhancing the models' predictive accuracy and efficiency. Notably, the tuning and training process for the embedding structure alternatives was executed with a 10% sample of the data with all unique activities of "Occupant_Activity" is included, providing a streamlined yet effective subset for model optimization. This preference is based on the limited computational capacity.

Tuning Process and Parameters for Embedding Structure Alternatives

To reach of best performance of the three distinct LSTM models, a uniform tuning process was applied. The method is the of Bayesian optimization with the Python-based Optuna tuning tool. This process involved the exploration of a diverse range of hyperparameters to present the optimal configuration for each model. After tuning process, each LSTM model was trained using their best parameters derived from this process. The parameters under consideration included as follows:

- the number of hidden layers (`num_hidden_layers`), i.e. ranging from 1 to 5
- the number of units in each hidden layer (`hidden_units`), i.e. ranging from 25 to 125
- various batch sizes (`batch_size`) of 24, 48, and 96
- learning rates (`learning_rate`) were varied among 0.001, 0.0005, and 0.0001
- embedding sizes (`embedding_size`) were tested at intervals from 25 to 100
- choice among different optimizers (`optimizer`), i.e. “adam”, “rmsprop”, “sgd”
- initializers (`initializer`) i.e. “glorot_uniform”, “he_normal”, “lecun_normal”
- activation function (`activation_function`) ie., “tanh” and “relu”
- the consideration of bidirectional layers (`bidirectional`) as binary choice

Results Of Training Process of All Alternative Embedding Structures With Its Tuning Versions

The LSTM models with different embedding structures adopted a distinct set of parameters during the tuning process. The first model simplifies its approach with four hidden layers, with a more diverse range of units (25 to 125). Due to its higher embedding capacity, it can efficiently process the less complex data structure.

In contrast, the second model uses five hidden layers but varies significantly in unit count (ranging from 50 to 125). Thus, it is indicating a different approach to balance between capturing temporal sequences and embedding categorical data.

A more complex architecture is adopted by the third model, as featuring five hidden layers. Each of these layers contains a high number of units, i.e. ranging 100 or 125. This design choice suggests that the model is intended to identify intricate patterns within the data as this model do not consist of any embedding structure.

All models employ the "relu" activation function which is capable for its effectiveness in non-linear transformations and a "bidirectional" structure that is offering higher power for capturing the data sequence. The strategic design choices made for each model are highlighted by the observed variations in layer complexity, unit distribution, and model depth. These variations facilitate the optimal processing of their respective data types within the context of occupancy modeling.

4.6.1 The First Model (ie. Without Embedding Structure)

The initial configuration of the model includes four hidden layers with a distinctive distribution of hidden units: 50 units for the first layer, 25 for the second, 75 for the third, and 125 for the fourth. It operates with a batch size of 96 and an embedding size of 75. The model utilizes the "glorot_uniform" and "lecun_normal" initializers for different aspects of its architecture. Optimization is managed by the "adam" optimizer. The "relu" function serves as the activation method, and the LSTM is bidirectional as enhancing its ability to process data.

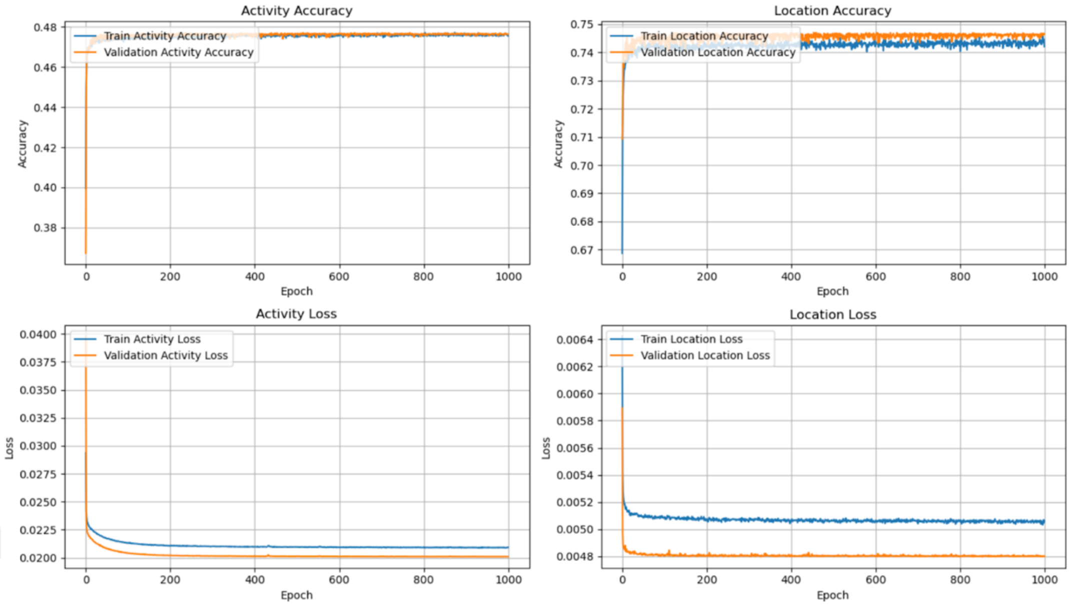

Figure 4.24. Performance of the first model

Figure 4.24 for this model show that while the training and validation accuracy for activity and location classification achieve a steady state, they do so at a modest level compared to the other models. This suggests that while the model is stable and consistent, its ability to learn and generalize complex patterns from the data is highly limited without the aid of categorical embeddings. The loss curves are showing a gradual decline and leveling out at a lower plateau. In short, the first model's architecture demonstrates how different architectural choices can yield varied learning outcomes based on the feature set provided.

4.6.2 The Second Model (ie. Embedding with Partial Categorical Columns)

The second model demonstrates a better performance but still cannot entirely capture the dataset's non-linearity. This model employs a deep architecture of five hidden layers, with a strategic variation in the number of units (`hidden_units_l0: 100`, `hidden_units_l1: 125`, `hidden_units_l2: 50`, `hidden_units_l3: 50`, `hidden_units_l4: 125). The large `batch_size` of 96 and a substantial `embedding_size` of 100 suggest a capacity for handling broad data features, while maintaining the `adam` optimizer, `lecun_normal` initializer, `relu` activation, and bidirectional LSTM layers as same for all the alternatives.

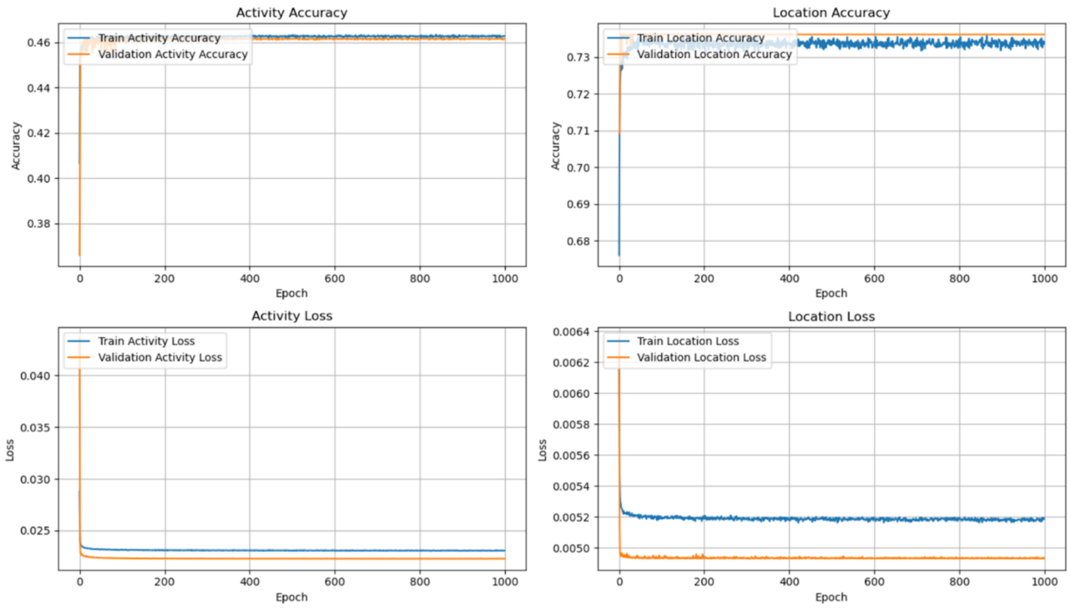

Figure 4.25. Performance of the second model

Figure 4.25 indicate average training consistency. This is resulted by the training and validation accuracy curves that remain closely aligned across 1000 epochs. While the activity accuracy is moderately high. As it plateaus early, it is suggesting a certain level of difficulty in learning more complex patterns within the data. Similarly, the location accuracy shows stability, yet it does not reach the same high levels as the third model. The loss curves for both activity and location flatten out at a higher level than desired. This could be indicative of the model's challenge in reducing error when faced with the intrinsic non-linearities of the dataset.

4.6.3 The Third Model (i.e. Embedding with Full Categorical Columns)

The third model is the main model of the imputation process, and it presents a high performance with its wide architecture. This model is designed for complexity as operating with a configuration of five hidden layers, uniformity in hidden units across the initial and final layers, and a heightened number in the intermediate layer *(`hidden_units_l0: 100`, `hidden_units_l1: 100`, `hidden_units_l2: 125`, `hidden_units_l3: 100`, `hidden_units_l4: 100`)*. The moderately sized `batch_size` of 48, combined with a `learning_rate` of 0.001, an `embedding_size` of 25, the use of `adam` for optimization, `lecun_normal` initializer, `relu` activation, and a `bidirectional` approach. Lower value selection for batch and embedding size shows that the model has better capacity to learn efficiently from the dataset.

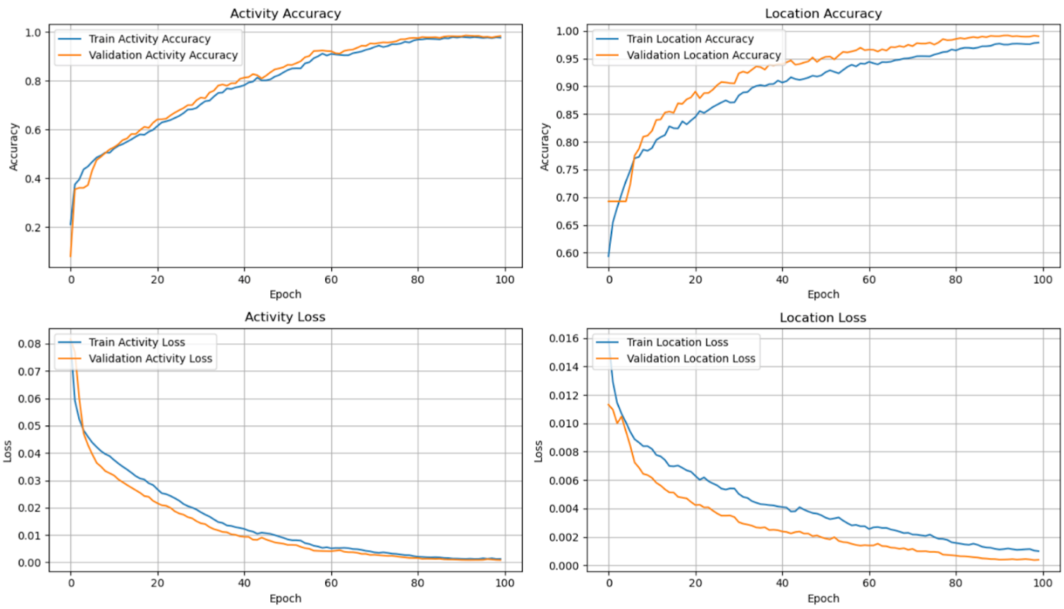

Figure 4.26. Performance of the third model

The performance reinforces the efficacy of the model, as shown in Figure 4.26. The training and validation accuracy for both activity and location classification demonstrate a consistent and converging improvement over epochs. with the validation curves closely shadowing the training ones. This is a sign of good generalization. Particularly, the validation accuracy for location reaches near-perfect levels and the model's capacity can discern occupant location with high precision. Similarly, the loss graphs exhibit a steady decline as plateauing as the model achieves a state of minimal error. The third model's architecture is thereby validated as not only robust but also adept at capturing the complex interplay of factors influencing daily occupant activities and locations within urban residential settings.

4.7 Experiment 3b: EM vs Various RNNs

LSTM and RNNs have been recognized as the precursors to state-of-the-art neural network models for processing sequential data, e.g. time series, text, and audio. While RNNs offer a more fundamental approach, they are concluding the problem of exploding gradient. The “gated” designs of LSTMs are employed to mitigate this issue. The chapter examine the distinctive features of these recurrent layers and evaluate their performance in the specialized context of occupancy modeling as similarly conducted in the previous sub-chapter. A comparative analysis is presented as focusing on two principal recurrent neural network architectures in terms of sequential deep learning applications for TUS temporal dataset. This analysis is characterized by a novel integration of temporal sequence datasets with non-temporal occupant demographic data. Each architecture is configured with an identical embedding structure by incorporating full categorical columns, periodical features, and detailed occupant demographics.

The Comparison Approach for different RNNs

Similar to the comparison of embedding structures, the recurrent neural network models were rigorously evaluated based on their training performance. This comparison utilized the best parameters identification through Bayesian optimization with same hyperparameters and same ranges. The validation process for training assessed both accuracy and loss metrics across model.

The training and tuning procedures for the recurrent model architectures were executed using a 10% sample of the data with all unique activities of “Occupant_Activity” included. Additionally, the architectural differences of each model were analyzed to present their respective capacities to handle the integrated dataset, which combines temporal sequences with non-temporal occupant demographics.

4.7.1 Results of Training Process for the First Model, LSTM

This LSTM model is distinguished by its utilization of both hidden states and cell states. It is a design that provides a solution to the vanishing gradient problem often encountered in standard RNNs. The architecture comprising five hidden layers with consistent unit allocation in the first and last layers, and a peak in the intermediary third layer *(hidden_units_l0: 100, hidden_units_l1: 100, hidden_units_l2: 125, hidden_units_l3: 100, hidden_units_l4: 100)*. This is built for the purpose of capturing complex patterns over extended sequences. A batch_size of 48, learning_rate of 0.001, embedding_size of 25, the adam optimizer, lecun_normal initializer, relu activation, and bidirectionality all contribute to the model's proficiency in learning from the dataset.

The LSTM's performance exhibits a remarkable and sustained improvement across epochs as demonstrated by the training and validation accuracy. The near-perfect validation accuracy for location highlights the LSTM's ability to accurately model occupant locations which is important aspect of understanding energy usage patterns in residential buildings.

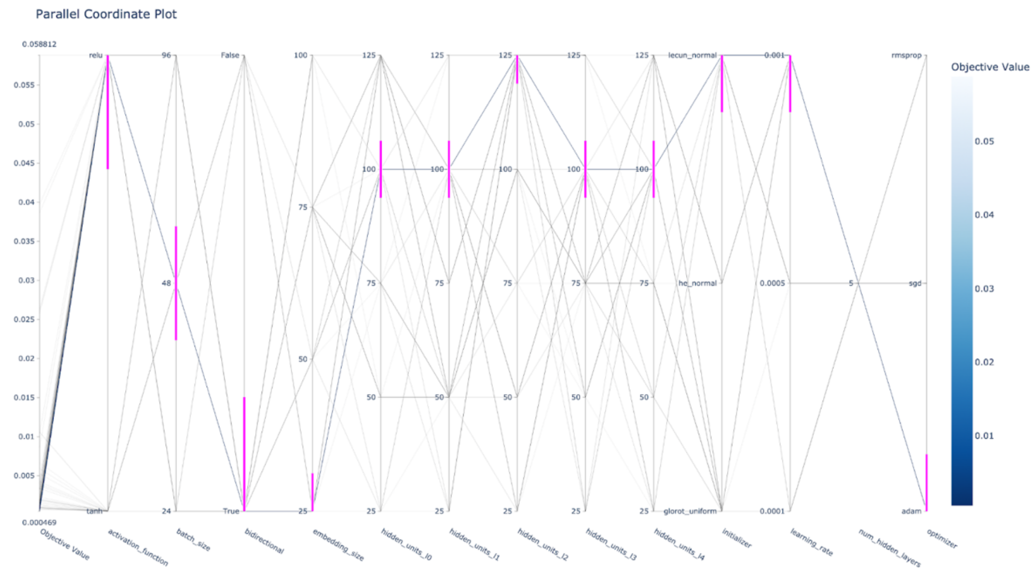

Figure 4.27. Hyper-Parameter Space visualization with Parallel Coordinate Plot

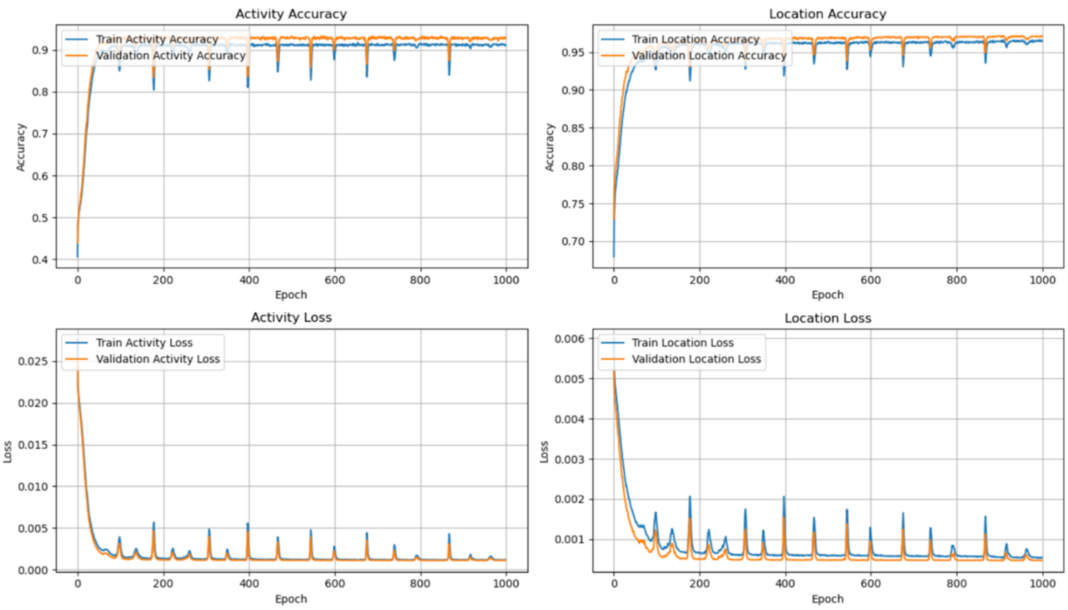

Figure 4.28. Performance of the first model (i.e. embedding with full categorical columns)

The declining loss curves underscore the LSTM's capability to minimize error, achieving optimal performance as shown in Figure 4.28. Unlike RNNs, the LSTM's unique architecture with cell states effectively overcomes the limitations associated with short-term memory. This is allowing it to retain information over longer periods and thus it captures the intricate dynamics of daily occupant activities and locations in urban residential environments.

4.7.2 Results of Training Process for the Second Model, LSTM

The RNN model is configured with a streamlined architecture of three hidden layers (hidden_units_l0: 125, hidden_units_l1: 100, hidden_units_l2: 50), and it was optimized through a meticulous tuning process. With a batch_size of 96, learning_rate of 0.001, embedding_size of 25, *rmsprop* optimizer, *glorot_uniform* initializer, *relu* activation function, and bidirectionality enabled, the model embarked on the training phase.

The training performance plots reveal that the RNN model demonstrates commendable learning dynamics., the RNN model's performance plateaus below the 90% threshold despite the impressive nature of its training trajectory. This is often associated with a higher level of generalization achieved by LSTM models. This indicates that while the RNN is capable of learning from the dataset, it struggles to encapsulate the full spectrum of variability inherent in the daily activity schedules.

Consequently, the results suggest a limitation in the RNN's ability to model complex patterns with the same efficacy as its LSTM counterpart, particularly in scenarios that demand the understanding of long-term dependencies and nuanced temporal sequences within occupant behavior data.

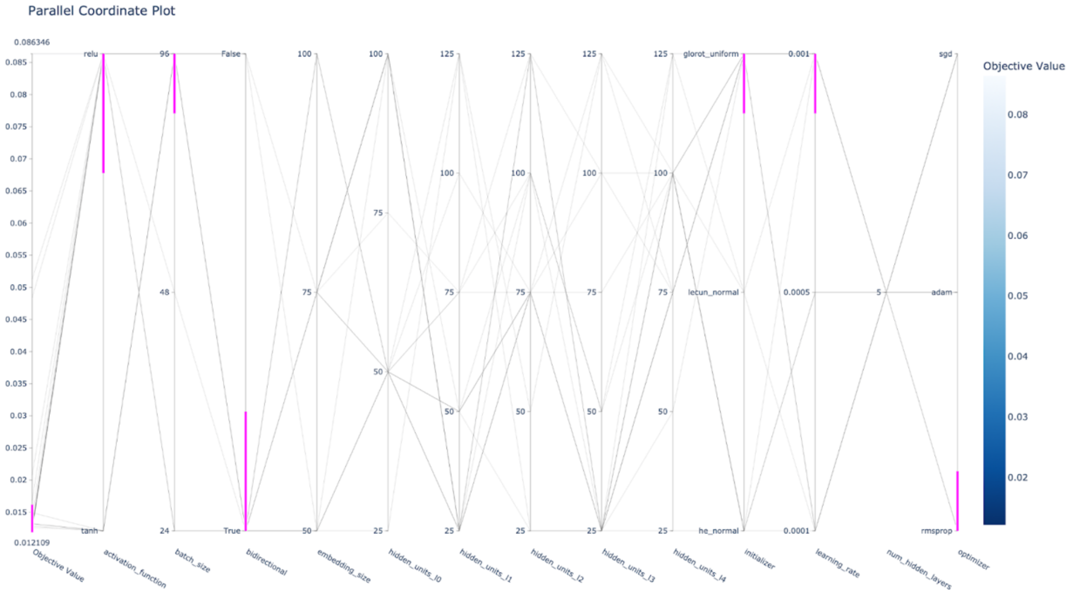

Figure 4.29. Hyper-Parameter Space visualization with Parallel Coordinate Plot

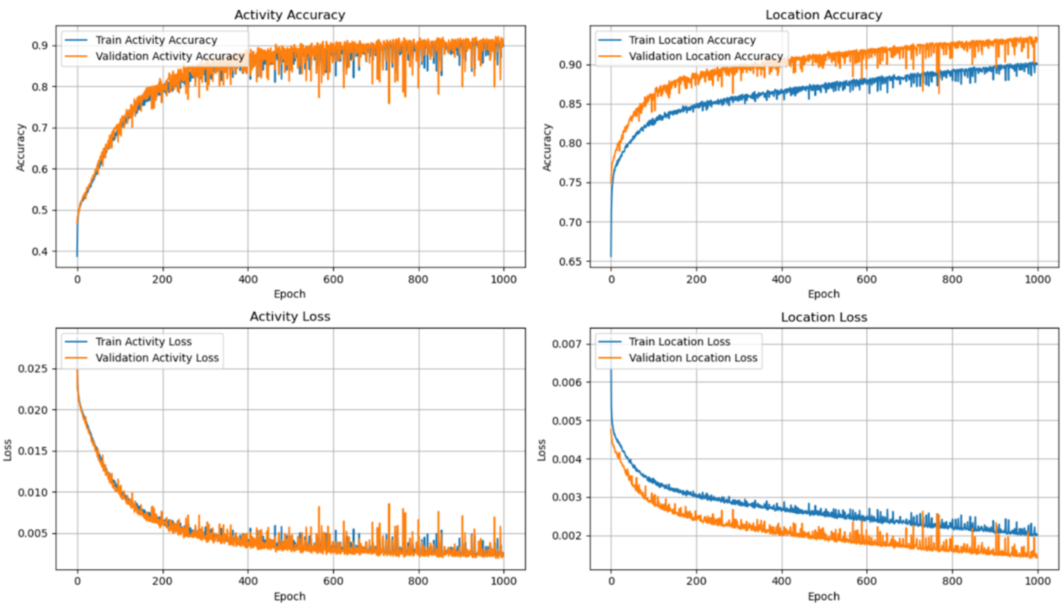

Figure 4.30. Performance of the second model (i.e. embedding with full categorical columns)

CHAPTER 5

CONCLUSION

A novel approach to residential building occupancy has been developed to address the discrepancy between expected and actual energy demand in urban residential buildings, which is mainly due to simplified or incorrect representations of occupancy. A multi-layered methodology was introduced as dependent occupancy modelling that combines non-sequential and sequential occupancy components using advanced data-driven techniques with deep learning algorithms without assumptions. This approach considers demographics, behaviors, interactions, and diversity inherent in the urban living environment. The approach uses time series data from population statistics to enable a deeper understanding of occupant behavior and its impact on energy demand with higher temporal resolution. This significant step not only elevates the accuracy of energy predictions in urban neighborhoods, but also opens the door to a commitment to a sustainable urban future.

5.1 Revisiting the Research Questions

Occupancy and Data-driven Methodology

Main Research Question: *How can a data-driven pipeline, integrating both temporal and non-temporal population statistics, be developed to enhance occupancy modeling better representation for urban residential buildings?*

When considering how a data-driven pipeline can improve occupancy modeling for urban residential buildings, it is important to clarify the relationship between the method and the data. The model in question was created based solely on population statistics, without relying on assumptions, rule-based approaches, private data or just energy consumption data. This foundation in population statistics allows for a generalized representation of user behavior and provides a comprehensive, systematic approach that avoids focusing on outliers. By integrating both temporal and non-temporal features, the model benefits from a comprehensive input structure that deep learning algorithms can process effectively. These algorithms are able to identify complex patterns in the data and capture key elements of user behavior across time and space, ultimately leading to a more accurate and scalable occupancy model for urban living environments.

Building on the foundations of population statistics, model development begins with the construction of a pipeline that seamlessly integrates both non-temporal and temporal information. Deep learning algorithms, particularly those developed for complex data patterns, are used to process and break down these data sets to provide a more accurate and data-driven representation of user behavior. This close connection between data and methodology ensures that the model reflects real-world behavior without relying on assumptions or rule-based systems.

The proposed method dives deeper into daily activities and presence patterns, translating them into a comprehensive view of occupancy across an entire year. Leveraging LSTM networks, which excel in handling sequential data, the model classifies occupant activity with a high degree of temporal resolution. This approach surpasses traditional methods, such as agent-based or stochastic modeling, by capturing the dynamic nature of occupant behavior in 30-minute intervals, encompassing activity, presence, and energy use schedules. This fine-grained temporal resolution is key to accurately modeling the complex, year-round occupancy patterns in residential buildings.

Essentials of Occupancy

Sub-Research Question 1: *What are the key elements of effective occupancy modeling, and how does incorporating various factors—such as lifestyle, work* *schedules, personal preferences, and socioeconomic and cultural influences—into a comprehensive model enhance its accuracy?*

Occupancy modeling depends on the thorough integration of key elements: individual (ie. occupant) and household level demographics, timing of presence in the residential, behavioral patterns, interpersonal dynamics, and energy use patterns. When these aspects are combined, a detailed insight into occupants' behavior can be achieved because all the components are dependent on each other because it is beneficial to communicate each component to each other as all they represent the one side of the occupants. Details about occupants form the basis of their profiles, i.e. age and economic background. Information about when and how occupants are present and active in a building is crucial for fine-tuning how the building operates. Activities as appliance usage and thermostat settings have a direct impact on energy demand and usage patterns.

Dependent occupancy modeling necessitates the integration of both non-temporal and temporal data to achieve a comprehensive and realistic depiction of occupant behavior. Non-temporal features, such as demographics from sources like CENSUS or Time-Use Surveys (TUS), offer crucial insights into factors like lifestyle, socioeconomic status, and cultural influences, all of which directly shape temporal patterns such as work schedules, routines, and personal preferences. These non-temporal datasets not only provide context for interpreting temporal data, but temporal behaviors also add depth and clarity to demographic trends. This interdependence forms an interconnected model, where both data types enrich each other, resulting in more accurate and dynamic representations of occupant activity and energy usage in urban residential areas. Additionally, the complex social interactions within households further influence collective decisions regarding daily activities and energy use. By incorporating personal characteristics and household interactions into the modeling process, this approach captures the multifaceted nature of occupancy, moving beyond traditional assumption-based models to deliver more precise and nuanced predictions of occupant behavior. In this way, the model is able to account for both broad demographic trends and the intricate daily routines of individuals, providing a complete and realistic solution that enhances the overall accuracy of energy modeling.

The permutation feature importance results reveal that occupant demographic features (eg. particularly age classes, gender, and employment status), have a significantly higher impact on occupancy modeling performance compared to temporal features such as "week or weekend" and "months or seasons." This highlights that non-temporal demographic data sourced from CENSUS and TUS datasets is more influential in shaping occupant behavior models than time-based variables alone.

Experiment with embedding differentiation tested the effectiveness of embedding non-temporal demographic features into the temporal flow of a deep learning model. This approach aimed to combine both temporal and non-temporal data to improve occupancy modeling accuracy. The experiment results showed that incorporating embedded demographics led to a remarkable improvement, increasing the model's accuracy by at least 50%. This enhancement was particularly evident when predicting target temporal features such as "occupant activity" and "location of the activity", demonstrating that demographic data significantly contributes to the overall accuracy of the model, in conjunction with temporal variables.

Occupancy and Dataset

Sub-Research Question 2: *What methods and privacy-preserving strategies can be used to balance the need for effective data collection in occupancy modeling while addressing the trade-off between the privacy risks datasets and the generalized representation capacity offered by datasets?*

Private datasets often come with significant privacy concerns, as the detailed, individualized data they contain can potentially expose sensitive information about occupants. Privacy-preserving strategies such as anonymization or differential privacy are necessary, but these measures can reduce data granularity and accuracy. Public datasets, on the other hand, are typically collected with permission and are more likely to be large-scale, offering a generalized representation of occupant behavior while minimizing privacy risks. These public datasets provide a valuable alternative for occupancy modeling, as their large scope can capture broader trends without compromising individual privacy.

For Urban Building Energy Modeling (UBEM), generalized occupancy representation is often more suitable, particularly when dealing with large-scale building performance simulations. Outlier values in private datasets can skew results, leading to inaccurate modeling outcomes. Public datasets, with their extensive coverage, help mitigate this issue by offering a broader, more representative sample. However, most public datasets were not originally collected with UBEM in mind, as they tend to focus on social or economic aspects. This makes data preprocessing essential for converting the dataset into a useful form for occupancy modeling, requiring careful attention to feature selection, categorization, and restructuring of the data to align with the goals of the UBEM process.

Relying on public datasets, such as those provided by formal institutions like ISTAT (The Italian National Institute of Statistics), is crucial to avoid assumption-based approaches. Publicly available datasets, such as population statistics, Time-Use Surveys, and building layouts, are particularly suitable as they offer in-depth demographic details essential for generating occupancy models with both sequential and non-sequential data, while also addressing privacy concerns. These datasets are collected with rigorous privacy standards, ensuring that personal data is anonymized and securely handled, making them a reliable and ethical choice for research.

Sub-Research Question 3: *What types of information are essential for occupancy datasets to achieve comprehensive occupancy modeling, and what levels of resolution are required regarding data types?*

By incorporating both temporal and non-temporal data, occupancy models can evolve from static to dynamic representations, capturing the full complexity of occupant behavior over time. Temporal data, such as daily or seasonal activity patterns, and non-temporal data, such as demographic details and household characteristics, are both essential for building accurate and robust models. The importance of preprocessing and integrating these diverse data types cannot be overstated, as it requires a complex process of thorough cleaning, filtering, and refining to ensure data quality and reliability. These approaches provide a comprehensive perspective on how occupants interact in residential spaces, greatly enhancing the precision and development of occupancy modeling in urban settings.

Sub-Research Question 4: *How can occupancy modeling be modified for global applicability, and what types of datasets should be chosen to support this objective?*

The methodology developed in this thesis can be adapted for global application by leveraging the standardized data collection approaches employed in initiatives like the Harmonised European Time Use Surveys (HETUS). HETUS, conducted across multiple European countries, follows consistent methodological guidelines that ensure a high degree of comparability between datasets from different regions. This standardization allows for the proposed occupancy modeling techniques to be applied effectively in any country that adopts similar data collection methods. The availability of microdata from HETUS 2010, which includes detailed temporal and non-temporal information from 17 European countries, demonstrates the feasibility of extending this methodology to diverse geographic contexts. As more countries participate in HETUS, particularly with the ongoing 2020 round, the global applicability of this approach is further enhanced. By utilizing these harmonized datasets, the methodology accommodated a methodology by considering the variability of data across regions, enabling accurate and scalable applications in urban building energy modeling worldwide.

Occupancy Modeling

Sub-Research Question 5: *What are the limitations of cluster-based occupancy modeling in accurately reflecting the complex nature of occupant behavior?*

The clustering for occupancy modeling uses demographic archetype profiling, however, it is relatively simple to implement and offers interpretability, it falls short in generating realistic, occupant-specific activity patterns. Clustering models focus on demographic data without fully accounting for the broader context of household dynamics and inter-occupant interactions, leading to incomplete modeling of nuanced daily routines. This approach primarily centers on demographics rather than the detailed interactions within the household. Furthermore, the clustering method struggles to capture seasonal or weekly variations accurately due to its limited evaluation scope, focusing solely on the number of clusters rather than individual-level details. Thus, the limitations of the clustering-based approach highlight the need for a more comprehensive method that incorporates both family composition and the intricate dynamics of daily interactions within a household.

Sub-Research Question 6: *What are the limitations of assumption-based occupancy modeling in accurately reflecting the stochastic nature of occupant behavior?*

Stochastic occupancy models, including techniques such as Markov chains, face significant limitations in accurately capturing the stochastic nature of user behavior. These models are often based on binary states and exhibit a memoryless property, i.e. they only consider the current state without considering the historical context. This limits their predictive power. They are also very sensitive to the initial conditions, which can lead to inaccuracies, especially if these conditions are uncertain or fluctuate widely. The fact that they are based on assumptions rather than actual data further exacerbates these limitations. The assumption-based essence of these types of models introduce uncertainty and do not reflect the complex and dynamic nature of real occupancy patterns. Consequently, the accuracy and reliability of the model are reduced.

Sub-Research Question 7: *What are the implications of modeling occupancy at the household level versus the individual occupant level for accuracy of occupancy modeling?*

A significant limitation in current occupancy modeling is its reliance on household-based approaches to determine presence schedules, which often overlook the individual variability and unique patterns of each occupant within a household. This approach tends to miss the intricate interactions and diverse behaviors of individual occupants, leading to less precise occupancy representations. In contrast, adopting an occupant-level modeling approach improves the granularity of occupancy analysis by treating each occupant as an independent agent capable of performing various tasks at different times. The proposed method (ie. End-to-End Deep Learning) captured the complex dynamics within a household that household-level models might miss by learning the individual-level actions and behaviors. Additionally, analyzing energy demand at the occupant level enables more efficient energy use and facilitates the development of personalized energy conservation strategies, providing valuable insights into demand patterns and leading to targeted energy-saving measures aligned with individual routines and preferences.

Sub-Research Question 8: *What are the key challenges in applying deep learning algorithms to the combination of time-series and time-series data in the context of occupancy modeling, and how can they be addressed?*

A key challenge in applying deep learning algorithms, such as LSTM models, to the combination of time series and non-temporal data in occupancy modeling lies in the effective integration of non-temporal categorical variables. The embedding of these non-temporal characteristics, such as demographic characteristics of residents and households, is crucial for synchronization with temporal data. By using dense layers for embedding, the model transforms sparse, high-dimensional categorical data into a denser, continuous representation and can thus recognize more complex patterns in the data. This integrated approach improves the learning performance of end-to-end LSTM models and makes them more effective for tasks involving multi-categorical and binary classifications. However, this advanced methodology also leads to higher computational complexity and cost, which poses additional challenges that need to be addressed to optimize the efficiency and scalability of the model.

One of the key challenges in applying deep learning algorithms to the combination of time-series and non-temporal data in occupancy modeling is balancing accuracy expectations between fields like social sciences and engineering. In social sciences, accuracy rates around 60% can be often acceptable due to the complexity of human behavior and social phenomena, while in engineering and natural sciences, higher accuracy—typically around 90%—is expected. This difference in standards can lead to criticism when deep learning models are applied to interdisciplinary contexts like occupancy modeling. In this study, accuracy metrics are used primarily to evaluate the learning performance of the models, particularly in terms of preventing overfitting or underfitting. The aim is to ensure that the models are generalizable and robust rather than focusing solely on achieving high accuracy numbers and capturing the complex dynamics of occupant behavior while maintaining practical relevance for both social and natural sciences.

Sub-Research Question 9: *How can deep learning techniques model the complex interactions between occupant profiles and household activities, and how do different modeling approaches compare in capturing this variability in residential UBEM?*

Conventional methods typically address components like presence and activity separately, failing to account for their intricate interrelationships. Advanced modeling techniques as deep learning methodology, particularly those capable of analyzing both sequential and non-sequential data concurrently, provide a more comprehensive and precise representation of how occupant behavior impacts energy demand. These models can effectively capture the dynamic interactions between occupants and their use of residential buildings and appliances, acknowledging that energy consumption is influenced by the interplay among occupants and their environment. Consequently, by integrating these multifaceted dependencies, the proposed deep learning approach significantly improves the accuracy of occupancy modeling for UBEM implementation, offering valuable insights into the daily routines and preferences of building occupants.

Occupancy modeling and UBEM implementation

Sub-Research Question 10: *What is the impact of using augmented datasets with seasonal and weekly variations on the precision of machine learning models in predicting occupant behavior?*

Occupants' behaviors fluctuate throughout the seasons due to changing weather conditions, which in turn affect their daily actions and energy consumption patterns. In the context of Urban Building Energy Modeling (UBEM), these seasonal changes directly influence indoor energy usage, particularly in terms of heating, cooling, and electrical demand. The choice of temporal resolution must balance the desired level of detail with the associated computational complexity. This approach, particularly when combined with longitudinal data collection over seasons and years, provides valuable insights into the evolving nature of occupancy patterns, ultimately improving the precision of machine learning models in predicting occupant behavior within residential buildings.

5.2 Limitations and Challenges

This research faces several limitations and challenges in terms of data collection, pre-processing, and modelling steps. Problems arise from the availability and compatibility of data sets, the accurate conversion of different data types in pre-processed input dataset and the complexity of modeling advanced algorithms.

5.2.1 Data Collection

HETUS and TUS from Türkiye: TUS data from Türkiye is covering residential aspects across 81 cities, however it is limited in scope of publicly available only at the city scale and lacks detailed occupant characteristics. In addition, this dataset does not comply with the HETUS standards which poses challenges for global implementation of the proposed LSTM model. As Turkey is a candidate for HETUS, a future adaptation could improve the applicability of the system. However, as the raw TUS data is currently not available and only aggregated statistics are available, alternative datasets that are more suitable for the accurate occupancy modeling for the UBEM implementation in Turkey.

TUS weekly and Employment status: TUS dataset consist of three componenets, TUS individual, TUS daily and TUS weekly. The use of weekly TUS data in occupancy modeling is limited as it composed of only for employed occupants. There is a lack of comprehensive data for other employment status, e.g. house-person, seniors, children etc. This limits the ability to create detailed models that differentiate between individual days of the week. Thus, TUS weekly is excluded from the modeling process.

5.2.2 Preprocessing

Occupant demographic selection: The selection of most effective parameters for embeddings in LSTM requires a focused impact analysis. However, the main challenge is to bridge the gap between sequential target variables, i.e. “occupant_activity” and “location” and non-sequential demographics i.e. “Region”, “Age Classes” and “Education Degree”. Thus, feature engineering (FE) is used to convert the sequential data into non-sequential format. This process is an integral part of the analysis, but involves inherent assumptions because the techniques selected for FE do not promise for fully capturing the details of the sequential patterns. A number of engineered features are created to deepen the granularity of the analysis:

- aggregate metrics ('distinct activities'),
- temporal patterns ('start_of_day', 'max_consecutive'),
- changing points ('activity changes')
- transformation techniques ('seasonal variation', 'activity entropy')
- location features as 'Total_Hours_Outside' and 'Total_Hours_Inside'

Despite the benefits of FE in correlating demographic characteristics with daily occupancy patterns, it is important to acknowledge the potential limitations of this approach. It is possibly oversimplifying the dynamic nature of the original sequential data may be oversimplified.

Merging for CENSUS and TUS: The integration TUS and CENSUS datasets for household population modeling is challenging due to discrepancies of the categories and contents of the columns. The problems include inconsistencies in the identification of households and discrepancies in categorical definitions, i.e. the number of family members and age categories. Despite these differences, the CENSUS dataset is invaluable as it provides comprehensive insights into linking residential details with household which is that are crucial for the UBEM simulations. Consequently, it is decided to use the CENSUS data separately to improve occupancy modeling. This approach avoids potential accuracy issues in the TUS data augmentation that could arise from attempting to combine these different datasets using advanced algorithms.

5.2.3 Modeling

LSTM with embedding structure: When developing sequential LSTM model, the embedding of non-temporal categorical variables played a crucial role. By using dense layers for embedding, the model effectively transforms sparse, high-dimensional categorical data into a denser, lower and continuous representation. In addition, embedding layers offer advantages over traditional encoding methods (i.e. one-hot coding) as they can enable the model to recognise more complex patterns within the categorical data. The integrated architecture of embeddings and LSTM proves to be a robust framework for tasks involving multicategorical and binary classifications. However, this advanced approach also increases computational complexity and cost.

Parallelization and computation efficiency of LSTM models: Research into different neural network architectures brings different advantages. While LSTM are a cornerstone in processing sequential data, however, there are inherent aspects while transformer-based time series prediction algorithms can offer complementary strengths. Transformers distinguish themselves in their ability to process entire sequences in parallel which they are resulting in higher training efficiency. In contrast, LSTMs framework is hard to parallelise.

Transformers show a strong ability to capture wide-ranging dependencies and contextual nuances through their self-attention mechanisms, especially beneficial for annual representation.

In terms of scalability and handling large-scale amounts of data tasks, transformers have proven to be more efficient as they scale better with increasing data and computational resources.

In terms of memory efficiency, especially when processing long sequences, Transformers manage resources more effectively through their attention-based mechanisms than the recurrent LSTMs, which might be less efficient in similar contexts.

However, as transformers are a newer area of technology primarily explored within language tasks, they are not yet widely implemented across various domains and require substantial computational power, despite offering options for parallelization and memory efficiency.

Recognizing these both positive and negative characteristics, this study intends to test the application of transformers in occupancy modeling to achieve potential improvements in model performance and accuracy.

5.3 Main Contributions

This thesis introduces a new approach that uniquely combines empirical data gathering, data processing and data-driven techniques for occupancy modeling. The use of deep learning on time-series data from population statistics offers detailed insights into the effect of occupant behavior on energy demand, marking a significant departure from conventional urban building energy modeling methods.

This research proposes:

- A dependent occupancy model that simultaneously considers occupant profiles, schedules, and interactions with the environment, going beyond methods with isolated parameters to illuminate the complex interplay between occupant behavior and energy demand in residential buildings.
- An advanced model to analyze occupant demographics along with hourly, occupants' weekly patterns (shaped by seasonal and weekly diversifications) and categorized energy use (lighting, appliances, HVAC) with measurable metrics.
- Openining a road to establish a national occupant-centric database for a solid foundation for occupancy models, setting new standards for future research,

and provide scalability in residential occupancy modeling potentially applicable to HETUS countries.

- A novel data-driven method that employs Long Short-Term Memory networks for granular presence and activity classification and provides a dynamic behavioral model that can narrow the performance gap and improve the understanding of energy patterns for UBEM.
- A multi-task modeling approach that simultaneously classifies presence and activity outputs represents a significant advance in accuracy, an achievement previously unattained in the literature, which typically focuses on either demographic data or only attendance schedule outputs.
- An LSTM-based model represents a departure from traditional clustering and Markov models that rely on assumption-laden characteristics. Instead, real population statistics are processed directly, minimizing assumptions, and providing a more realistic occupancy model.
- A guide for the integration of occupancy components in UBEM simulations, detailing how to convert and implement elements such as households (HHs), occupants (OCCs), schedules (SCHs) and activities (ACTs) in terms of operational and energy schedules (ENERs), alongside technical steps for the integration of these components in BEM using the EnergyPlus (E+) framework for the Italian context.

5.4 Further Works

This chapter explores potential advances, from the integration of advanced sequential models to the creation of lighting and equipment schedules, all aimed at improving model accuracy, broadening applicability (including the Italian context) and ultimately enabling more accurate UBEM for efficient urban energy planning.

First, there is an opportunity to investigate the application of various advanced sequential classification models, including transformers, to improve occupancy modeling. In addition, the creation of weekly TUS data for employed individuals could extend the data representation from three to seven days, providing a more comprehensive insight into weekly behavioral insight.

Another possibility is to extend the TUS dataset, even though the household and occupant related non-sequential characteristics are present in both the TUS and CENSUS datasets; the ID columns do not match. Therefore, merging the two datasets requires advanced approaches which it is excluded from this study. For further work, the two datasets will be merged and then the total number of households for the whole of Italy will be increased.

Specifically in the Italian context, future work could focus on creating lighting and equipment density schedules based on daily activity schedules and moving beyond binary usage models. This advance would significantly improve the granularity and accuracy of occupancy modeling when analyzing the energy demand of residential buildings.

Finally, an important future step could be the implementation of the UBEM at district level in different Italian regions. This would present the impact of developed occupancy information within the proposed model, which could lead to more localized and accurate energy demand predictions and urban planning strategies.

REFERENCES

Abdallah, F., Basurra, S., & Gaber, M. M. (2017). A hybrid agent-based and probabilistic model for fine-grained behavioural energy waste simulation. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2017-November*. https://doi.org/10.1109/ICTAI.2017.00152

Aerts, D., Minnen, J., Glorieux, I., Wouters, I., & Descamps, F. (2014). A method for the identification and modelling of realistic domestic occupancy sequences for building energy demand simulations and peer comparison. *Building and Environment*, *75*. https://doi.org/10.1016/j.buildenv.2014.01.021

Ahmed, K., Kurnitski, J., & Olesen, B. (2017). Data for occupancy internal heat gain calculation in main building categories. *Data in Brief*, *15*. https://doi.org/10.1016/j.dib.2017.10.036

Ahmed, O., Sezer, N., Ouf, M., Wang, L. (Leon), & Hassan, I. G. (2023a). State-of-the-art review of occupant behavior modeling and implementation in building performance simulation. *Renewable and Sustainable Energy Reviews*, *185*, 113558. https://doi.org/https://doi.org/10.1016/j.rser.2023.113558

Ahmed, O., Sezer, N., Ouf, M., Wang, L. (Leon), & Hassan, I. G. (2023b). State-of-the-art review of occupant behavior modeling and implementation in building performance simulation. In *Renewable and Sustainable Energy Reviews* (Vol. 185). https://doi.org/10.1016/j.rser.2023.113558

Ainsworth, B. E., Haskell, W. L., Herrmann, S. D., Meckes, N., Bassett, D. R., Tudor-Locke, C., Greer, J. L., Vezina, J., Whitt-Glover, M. C., & Leon, A. S. (2011). 2011 compendium of physical activities: A second update of codes and MET values. In *Medicine and Science in Sports and Exercise* (Vol. 43, Issue 8). https://doi.org/10.1249/MSS.0b013e31821ece12

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3292500.3330701

Akkaya, K., Guvenc, I., Aygun, R., Pala, N., & Kadri, A. (2015). *IoT-based Occupancy Monitoring Techniques for Energy-Efficient Smart Buildings*. 58–63.

Aksoezen, M., Daniel, M., Hassler, U., & Kohler, N. (2015). Building age as an indicator for energy consumption. *Energy and Buildings*, *87*, 74–86. https://doi.org/10.1016/j.enbuild.2014.10.074

Alfalah, B., Shahrestani, M., & Shao, L. (2023). Developing a Hidden Markov model for occupancy prediction in high-density higher education buildings. *Journal of Building Engineering*, *73*. https://doi.org/10.1016/j.jobe.2023.106795

Ali, U., Shamsi, M. H., Hoare, C., Mangina, E., & O'Donnell, J. (2021). Review of urban building energy modeling (UBEM) approaches, methods and tools using qualitative and quantitative analysis. *Energy and Buildings*, *246*, 111073. https://doi.org/10.1016/j.enbuild.2021.111073

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10). https://doi.org/10.1093/bioinformatics/btq134

Anand, P., Cheong, D., Sekhar, C., Santamouris, M., & Kondepudi, S. (2019). Energy saving estimation for plug and lighting load using occupancy analysis. *Renewable Energy*, *143*. https://doi.org/10.1016/j.renene.2019.05.089

Anand, P., Deb, C., Yan, K., Yang, J., Cheong, D., & Sekhar, C. (2021). Occupancy-based energy consumption modelling using machine learning algorithms for institutional buildings. *Energy and Buildings*, *252*. https://doi.org/10.1016/j.enbuild.2021.111478

Andersen, P. D., Iversen, A., Madsen, H., & Rode, C. (2014). Dynamic modeling of presence of occupants using inhomogeneous Markov chains. *Energy and Buildings*, *69*. https://doi.org/10.1016/j.enbuild.2013.10.001

Andrews, C. J., Yi, D., Krogmann, U., Senick, J. A., & Wener, R. E. (2011). Designing buildings for real occupants: An agent-based approach. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *41*(6). https://doi.org/10.1109/TSMCA.2011.2116116

Armina, R., Mohd Zain, A., Ali, N. A., & Sallehuddin, R. (2017). A Review on Missing Value Estimation Using Imputation Algorithm. *Journal of Physics: Conference Series*, *892*(1). https://doi.org/10.1088/1742-6596/892/1/012004

Armstrong, M. M., Swinton, M. C., Ribberink, H., Beausoleil-Morrison, I., & Millette, J. (2009). Synthetically derived profiles for representing occupant-driven electric loads in Canadian housing. *Journal of Building Performance Simulation*, *2*(1). https://doi.org/10.1080/19401490802706653

Arnfield, J. (2023). Koppen climate classification. In *Encyclopedia Britannica*.

Ashouri, M., Haghighat, F., Fung, B. C. M., Lazrak, A., & Yoshino, H. (2018). Development of building energy saving advisory: A data mining approach. *Energy and Buildings*, *172*. https://doi.org/10.1016/j.enbuild.2018.04.052

ASHRAE. (2004a). *ASHRAE Standard 55-2004 – Thermal Comfort* (Vol. 2004). https://doi.org/10.1007/s11926-011-0203-9

ASHRAE. (2004b). *ASHRAE Standard 55-2004 – Thermal Comfort* (Vol. 2004). https://doi.org/10.1007/s11926-011-0203-9

ASHRAE. (2009). *ASHRAE climatic design conditions 2009/2013/2017*.

ASHRAE. (2013). *ASHRAE Standard 90.1-2013 – Energy Standard For Buildings Except Low-rise Residential Buildings*.

ASHRAE. (2016). Standard 90.1-2016. Energy Standard for Buildings Except Low-Rise Residential Buildings. *American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.* https://doi.org/http://dx.doi.org/10.1108/17506200710779521

Asim, N., Badiei, M., Mohammad, M., Razali, H., Rajabi, A., Chin Haw, L., & Jameelah Ghazali, M. (2022). Sustainability of Heating, Ventilation and Air-Conditioning (HVAC) Systems in Buildings—An Overview. *International Journal of Environmental Research and Public Health*, *19*(2). https://doi.org/10.3390/ijerph19021016

Azar, E., & C. Menassa, C. (2012). Agent-Based Modeling of Occupants and Their Impact on Energy Use in Commercial Buildings. In *Journal of Computing in Civil Engineering* (Vol. 26). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000158

Babaei, T., Abdi, H., Lim, C. P., & Nahavandi, S. (2015). A study and a directory of energy consumption data sets of buildings. *Energy and Buildings*, *94*. https://doi.org/10.1016/j.enbuild.2015.02.043

Baetens, R., & Saelens, D. (2016). Modelling uncertainty in district energy simulations by stochastic residential occupant behaviour. *Journal of Building Performance Simulation*. https://doi.org/10.1080/19401493.2015.1070203

Bakker, V., Molderink, A., Hurink, J. L., & Smit, G. J. M. (2008). Domestic heat demand prediction using neural networks. *Proceedings of 19th International Conference on Systems Engineering, ICSEng 2008*. https://doi.org/10.1109/ICSEng.2008.51

Ballarini, I., Corgnati, S. P., Corrado, V., & Talà, N. (2011). Definition of building typologies for energy investigations on residential sector by Tabula Iee-Project: Application to Italian case studies. *Roomvent*, 19–22.

Balvedi, B. F., Ghisi, E., & Lamberts, R. (2018). A review of occupant behaviour in residential buildings. *Energy and Buildings*, *174*, 495–505. https://doi.org/https://doi.org/10.1016/j.enbuild.2018.06.049

Bank, W. (2010). *Cities and climate change : an urgent agenda. Urban development series knowledge papers*.

Benezeth, Y., Laurent, H., Emile, B., & Rosenberger, C. (2011). Towards a sensor for detecting human presence and characterizing activity. *Energy and Buildings*. https://doi.org/10.1016/j.enbuild.2010.09.014

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*.

Bianchi, C., Zhang, L., Goldwasser, D., Parker, A., & Horsey, H. (2020). Modeling occupancy-driven building loads for large and diversified building stocks through the use of parametric schedules. *Applied Energy*, *276*. https://doi.org/10.1016/j.apenergy.2020.115470

Bianco Mauthe Degerfeld, F., Piro, M., De Luca, G., Ballarini, I., & Corrado, V. (2023). The application of EN ISO 52016-1 to assess building cost-optimal energy performance levels in Italy. *Energy Reports*, *10*. https://doi.org/10.1016/j.egyr.2023.08.014

Booten, C., Robertson, J., Christensen, D., Heaney, M., Brown, D., Norton, P., & Smith, C. (2017). *Residential Indoor Temperature Study*. https://doi.org/10.2172/1351449

Bosu, I., Mahmoud, H., & Hassan, H. (2023). Energy audit, techno-economic, and environmental assessment of integrating solar technologies for energy management in a university residential building: A case study. *Applied Energy*, *341*, 121141. https://doi.org/https://doi.org/10.1016/j.apenergy.2023.121141

Bottaccioli, L., Di Cataldo, S., Acquaviva, A., & Patti, E. (2019). Realistic Multi-Scale Modeling of Household Electricity Behaviors. *IEEE Access*, *7*. https://doi.org/10.1109/ACCESS.2018.2886201

Burak Gunay, H., O'Brien, W., & Beausoleil-Morrison, I. (2015). Development of an occupancy learning algorithm for terminal heating and cooling units. *Building and Environment*, *93*(P2). https://doi.org/10.1016/j.buildenv.2015.06.009

Buso, T., Fabi, V., Andersen, R. K., & Corgnati, S. P. (2015). Occupant behaviour and robustness of building design. *Building and Environment*, *94*. https://doi.org/10.1016/j.buildenv.2015.11.003

Buttitta, G., Turner, W. J. N., Neu, O., & Finn, D. P. (2019). Development of occupancy-integrated archetypes: Use of data mining clustering techniques to embed occupant behaviour profiles in archetypes. *Energy and Buildings*, *198*. https://doi.org/10.1016/j.enbuild.2019.05.056

Büyükalaca, O., Bulut, H., & Yılmaz, T. (2001). Analysis of variable-base heating and cooling degree-days for turkey. *Applied Energy*, *69*(4). https://doi.org/10.1016/S0306-2619(01)00017-4

Carpino, C., Mora, D., Arcuri, N., & De Simone, M. (2017). Behavioral variables and occupancy patterns in the design and modeling of Nearly Zero Energy Buildings. *Building Simulation*, *10*(6), 875–888.

Chapman, J., Siebers, P. O., & Robinson, D. (2018). On the multi-agent stochastic simulation of occupants in buildings. *Journal of Building Performance Simulation*. https://doi.org/10.1080/19401493.2017.1417483

Chen, J., Adhikari, R., Wilson, E., Robertson, J., Fontanini, A., Polly, B., & Olawale, O. (2021). *Stochastic simulation of residential building occupant-driven energy use in a bottom-up model of the U.S. housing stock*.

Chen, S., Zhang, G., Xia, X., Chen, Y., Setunge, S., & Shi, L. (2021). The impacts of occupant behavior on building energy consumption: A review. *Sustainable Energy Technologies and Assessments*, *45*, 101212. https://doi.org/https://doi.org/10.1016/j.seta.2021.101212

Chen, Y., Hong, T., & Luo, X. (2018). An agent-based stochastic Occupancy Simulator. *Building Simulation*, *11*(1). https://doi.org/10.1007/s12273-017-0379-7

Chen, Z., & Jiang, C. (2018). Building occupancy modeling using generative adversarial network. *Energy and Buildings*, *174*. https://doi.org/10.1016/j.enbuild.2018.06.029

Comitato Termotecnico Italiano. (2011). UNI EN 15251. In *Management*.

Comitato Termotecnico Italiano. (2013). *UNI EN 11439*.

Compendium. (2011). *Conversion Table from TUS to Compendium, MET*. https://docs.google.com/spreadsheets/d/1YtIxjUSHuBgwA0m3dH47D3IOVN1v2iFyLo2KwyT_nCM/edit#gid=1877476144

Cooijmans, T., Ballas, N., Laurent, C., Gülçehre, Ç., & Courville, A. (2017). Recurrent batch normalization. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*.

Cortellari, M., Barbato, M., Talenti, A., Bionda, A., Carta, A., Ciampolini, R., Ciani, E., Crisà, A., Frattini, S., Lasagna, E., Marletta, D., Mastrangelo, S., Negro, A., Randi, E., Sarti, F. M., Sartore, S., Soglia, D., Liotta, L., Stella, A., … Crepaldi, P. (2021). The climatic and genetic heritage of Italian goat breeds with genomic SNP data. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-89900-2

Crawley, B. D. B., Pedersen, C. O., Lawrie, L. K., & Winkelmann, F. C. (2000). EnergyPlus : Energy Simulation Program. *ASHRAE*, 49–56.

Dabirian, S., Panchabikesan, K., & Eicker, U. (2021). Occupant-centric Modeling for Urban Building Energy Simulation: Approaches, Inputs, and Data Sources - A Review. *Energy and Buildings*, *257*, 111809. https://doi.org/10.1016/j.enbuild.2021.111809

Dabirian, S., Panchabikesan, K., & Eicker, U. (2022). Occupant-centric urban building energy modeling: Approaches, inputs, and data sources - A review. *Energy and Buildings*, *257*, 111809. https://doi.org/https://doi.org/10.1016/j.enbuild.2021.111809

Davis, J. A., & Nutter, D. W. (2010). Occupancy diversity factors for common university building types. *Energy and Buildings*, *42*(9). https://doi.org/10.1016/j.enbuild.2010.03.025

D'Oca, S., Fabi, V., Corgnati, S. P., & Andersen, R. K. (2014). Effect of thermostat and window opening occupant behavior models on energy use in homes. *Building Simulation*, *7*(6). https://doi.org/10.1007/s12273-014-0191-6

D'Oca, S., & Hong, T. (2015). Occupancy schedules learning process through a data mining framework. *Energy and Buildings*, *88*. https://doi.org/10.1016/j.enbuild.2014.11.065

DOE. (2013). ASHRAE, Energy standard for buildings except low-rise residential buildings. *ASHRAE Standard*, *4723*(90.1-2013 (I-P)).

Dong, B., Liu, Y., Fontenot, H., Ouf, M., Osman, M., Chong, A., Qin, S., Salim, F., Xue, H., Yan, D., Jin, Y., Han, M., Zhang, X., Azar, E., & Carlucci, S.

(2021). Occupant behavior modeling methods for resilient building design, operation and policy at urban scale: A review. *Applied Energy*, *293*. https://doi.org/10.1016/j.apenergy.2021.116856

Dong, B., Liu, Y., Mu, W., Jiang, Z., Pandey, P., Hong, T., Olesen, B., Lawrence, T., O'Neil, Z., Andrews, C., Azar, E., Bandurski, K., Bardhan, R., Bavaresco, M., Berger, C., Burry, J., Carlucci, S., Chvatal, K., De Simone, M., … Zhou, X. (2022). A Global Building Occupant Behavior Database. *Scientific Data*, *9*(1), 369. https://doi.org/10.1038/s41597-022-01475-3

Dong, B., Yan, D., Li, Z., Jin, Y., Feng, X., & Fontenot, H. (2018). Modeling occupancy and behavior for better building design and operation—A critical review. *Building Simulation*, *11*. https://doi.org/10.1007/s12273-018-0452-x

Duarte, C., Budwig, R., & Van Den Wymelenberg, K. (2015). Energy and demand implication of using recommended practice occupancy diversity factors compared to real occupancy data in whole building energy simulation. *Journal of Building Performance Simulation*, *8*(6). https://doi.org/10.1080/19401493.2014.966275

Ekwevugbe, T., Brown, N., Pakka, V., & Fan, D. (2013). Real-time building occupancy sensing using neural-network based sensor network. *IEEE International Conference on Digital Ecosystems and Technologies*. https://doi.org/10.1109/DEST.2013.6611339

El Kontar, R., & Rakha, T. (2018). Profiling Occupancy Patterns to Calibrate Urban Building Energy Models (UBEMs) Using Measured Data Clustering. *Technology Architecture and Design*, *2*(2), 206–217. https://doi.org/10.1080/24751448.2018.1497369

EPBD. (2021). *Questions and Answers on the revision of the Energy Performance of Buildings Directive*. https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_6686

Erickson, V. L., Carreira-Perpiñán, M. Á., & Cerpa, A. E. (2011). OBSERVE: Occupancy-based system for efficient reduction of HVAC energy. *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks, IPSN'11*.

Esquivel, R. M., Gómez-Vargas, I., Vázquez, J. A., & Salcedo, R. G. (2021). An introduction to Markov Chain Monte Carlo. *Boletin de Estadistica e Investigacion Operativa*, *37*(1). https://doi.org/10.2139/ssrn.1594971

Eurostat. (2018). *Activity coding list for harmonised European time use surveys (ACL 2018)*.

eurostat. (2018). *Harmonised European Time Use Surveys (Hetus)*. 2018 Publication Guidelines. https://ec.europa.eu/eurostat/web/time-use-surveys/overview

Eurostat. (2019). *Harmonised European Time Use Surveys (HETUS) 2018 Guidelines*.

Fabi, V., Andersen, R. K., & Corgnati, S. (2015). Verification of stochastic behavioural models of occupants' interactions with windows in residential buildings. *Building and Environment*, *94*(P1). https://doi.org/10.1016/j.buildenv.2015.08.016

Fabi, V., Andersen, R. V., Corgnati, S. P., & Olesen, B. W. (2013). A methodology for modelling energy-related human behaviour: Application to window opening behaviour in residential buildings. *Building Simulation*, *6*(4). https://doi.org/10.1007/s12273-013-0119-6

FEDERICA ZAGARELLA. (2019). *Estimating the buildings hourly energy demand for Smart Energy District planning* [Politecnico di Milano]. https://www.politesi.polimi.it/handle/10589/144831

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

Flett, G., & Kelly, N. (2016). An occupant-differentiated, higher-order Markov Chain method for prediction of domestic occupancy. *Energy and Buildings*, *125*. https://doi.org/10.1016/j.enbuild.2016.05.015

Gaetani, I., Hoes, P.-J., & Hensen, J. L. M. (2016). Occupant behavior in building energy simulation: Towards a fit-for-purpose modeling strategy. *Energy and Buildings*, *121*, 188–204. https://doi.org/https://doi.org/10.1016/j.enbuild.2016.03.038

Garbin, C., Zhu, X., & Marques, O. (2020). Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, *79*(19–20). https://doi.org/10.1007/s11042-019-08453-9

Garg, V., Mathur, J., Tetali, S., & Bhatia, A. (2017). *Building Energy Simulation: A Workbook Using DesignBuilderTM*. CRC Press.

Grewal, J. K., Krzywinski, M., & Altman, N. (2019). Markov models—Markov chains. In *Nature Methods* (Vol. 16, Issue 8). https://doi.org/10.1038/s41592-019-0476-x

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., Huth, A., Jepsen, J. U., Jørgensen, C., Mooij, W. M., Müller, B., Pe'er, G., Piou, C., Railsback, S. F., Robbins, A. M., … DeAngelis, D. L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, *198*(1–2). https://doi.org/10.1016/j.ecolmodel.2006.04.023

Gunay, B., O'brien, W., Bcausolcil-Morrison, I., D'Oca, S., & Corgnati, S. (2015). On modelling and simulation of occupant models. *14th International Conference of IBPSA - Building Simulation 2015, BS 2015, Conference Proceedings*. https://doi.org/10.26868/25222708.2015.2825

Gunay, H. B., O'Brien, W., & Beausoleil-Morrison, I. (2013). A critical review of observation studies, modeling, and simulation of adaptive occupant behaviors in offices. *Building and Environment*, *70*, 31–47. https://doi.org/https://doi.org/10.1016/j.buildenv.2013.07.020

Haldi, F., & Robinson, D. (2009). Interactions with window openings by office occupants. *Building and Environment*, *44*(12). https://doi.org/10.1016/j.buildenv.2009.03.025

Haldi, F., & Robinson, D. (2011). The impact of occupants' behaviour on building energy demand. *Journal of Building Performance Simulation*, *4*(4). https://doi.org/10.1080/19401493.2011.558213

Happle, G., Fonseca, J. A., & Schlueter, A. (2017). Effects of air infiltration modeling approaches in urban building energy demand forecasts. *Energy Procedia*, *122*. https://doi.org/10.1016/j.egypro.2017.07.323

Happle, G., Fonseca, J. A., & Schlueter, A. (2018). A review on occupant behavior in urban building energy models. In *Energy and Buildings* (Vol. 174). https://doi.org/10.1016/j.enbuild.2018.06.030

Happle, G., Fonseca, J. A., & Schlueter, A. (2020). Context-specific urban occupancy modeling using location-based services data. *Building and Environment*, *175*. https://doi.org/10.1016/j.buildenv.2020.106803

Heydarian, A., McIlvennie, C., & Arpan, L. (2020). What drives our behaviors in buildings? A review on occupant interactions with building systems from the lens of behavioral theories. *Building and Environment*, *179*.

Heydenreich, J., Schutz, Y., Melzer, K., & Kayser, B. (2019). Comparison of conventional and individualized 1-MET values for expressing maximum aerobic metabolic rate and habitual activity related energy expenditure. *Nutrients*, *11*(2). https://doi.org/10.3390/nu11020458

Hong, S. H., Lee, J. M., Moon, J. W., & Lee, K. H. (2018). Thermal comfort, energy and cost impacts of PMV control considering individual metabolic rate variations in residential building. *Energies*, *11*(7). https://doi.org/10.3390/en11071767

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

Hong, T., Yan, D., D'Oca, S., & fei Chen, C. (2017). Ten questions concerning occupant behavior in buildings: The big picture. *Building and Environment*, *114*, 518–530. https://doi.org/10.1016/j.buildenv.2016.12.006

Hsu, D. (2015). Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. *Applied Energy*, *160*, 153–163. https://doi.org/10.1016/j.apenergy.2015.08.126

Huchuk, B., Sanner, S., & O'Brien, W. (2019). Comparison of machine learning models for occupancy prediction in residential buildings using connected thermostat data. *Building and Environment*, *160*. https://doi.org/10.1016/j.buildenv.2019.106177

ISTAT. (2013a). *15th Population and housing census 2011*.

ISTAT. (2013b). *Working Data: Censuses Of Population And Housing 2011 (1% Sample)*. ISTAT. https://docs.google.com/spreadsheets/d/1Q7Ho__jd4wlEhbtJe_e-k-lapZs5XjJPkKBKtadk-gk/edit#gid=1620366827

ISTAT. (2018). *TIMES OF DAILY LIFE Work, work-life balance, gender equality and well-being*. https://www.istat.it/it/files/2019/05/summary-Times-of-daily-life.pdf

ISTAT. (2022). INFORMAZIONI SULLA RILEVAZIONE MULTISCOPO SULLE FAMIGLIE: USO DEL TEMPO. In *ISTAT*.

Januzaj, Y., Beqiri, E., & Luma, A. (2023). Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique. *International Journal of Online and Biomedical Engineering*, *19*(4). https://doi.org/10.3991/ijoe.v19i04.37059

Jones, G. L., & Qin, Q. (2022). Markov Chain Monte Carlo in Practice. *Annual Review of Statistics and Its Application*, *9*. https://doi.org/10.1146/annurev-statistics-040220-090158

Julien Walzberg, T., Merveille, N., & Cheriet, M. (2017). An agent-based model to evaluate smart homes sustainability potential. *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, *2017-October*. https://doi.org/10.1109/PIMRC.2017.8292682

Kjærgaard, M. B., Ardakanian, O., Carlucci, S., Dong, B., Firth, S. K., Gao, N., Huebner, G. M., Mahdavi, A., Rahaman, M. S., Salim, F. D., Sangogboye, F. C., Schwee, J. H., Wolosiuk, D., & Zhu, Y. (2020). Current practices and infrastructure for open data based research on occupant-centric design and operation of buildings. *Building and Environment*, *177*. https://doi.org/10.1016/j.buildenv.2020.106848

Kontokosta, C. E. (2018). Urban Informatics in the Science and Practice of Planning. *Journal of Planning Education and Research*. https://doi.org/10.1177/0739456X18793716

Kotireddy, R., Hoes, P. J., & Hensen, J. L. M. (2017). Simulation-based comparison of robustness assessment methods to identify robust low-energy building designs. *Building Simulation Conference Proceedings*, *1*. https://doi.org/10.26868/25222708.2017.240

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, *15*(3). https://doi.org/10.1127/0941-2948/2006/0130

Kumar, S. (2007). Fourth assessment report of the Intergovernmental Panel on Climate Change: Important observations and conclusions. *Current Science*, *92*(8), 1034.

Labeodan, T., Zeiler, W., Boxem, G., & Zhao, Y. (2015). Occupancy measurement in commercial office buildings for demand-driven control applications - A survey and detection system evaluation. In *Energy and Buildings*. https://doi.org/10.1016/j.enbuild.2015.02.028

Langevin, J., Wen, J., & Gurian, P. (2014). *Including Occupants in Building Performance Simulation: Integration of an Agent-Based Occupant Behavior Algorithm with EnergyPlus*. https://doi.org/10.13140/2.1.2551.4244

Langevin, J., Wen, J., & Gurian, P. L. (2015). Simulating the human-building interaction: Development and validation of an agent-based model of office occupant behaviors. *Building and Environment*, *88*. https://doi.org/10.1016/j.buildenv.2014.11.037

LBNL, L. B. N. L. (2009). *Input Output Reference. EnergyPlus*.

Lee, Y. S., & Malkawi, A. M. (2014). Simulating multiple occupant behaviors in buildings: An agent-based modeling approach. *Energy and Buildings*, *69*. https://doi.org/10.1016/j.enbuild.2013.11.020

Li, J., Yu, Z. (Jerry), Haghighat, F., & Zhang, G. (2019). Development and improvement of occupant behavior models towards realistic building performance simulation: A review. *Sustainable Cities and Society*, *50*, 101685. https://doi.org/https://doi.org/10.1016/j.scs.2019.101685

Liang, X., Hong, T., & Shen, G. Q. (2016). Occupancy data analytics and prediction: A case study. *Building and Environment*, *102*. https://doi.org/10.1016/j.buildenv.2016.03.027

Liao, C., Lin, Y., & Barooah, P. (2012). Agent-based and graphical modelling of building occupancy. *Journal of Building Performance Simulation*, *5*(1), 5–25. https://doi.org/10.1080/19401493.2010.531143

Liisberg, J., Møller, J. K., Bloem, H., Cipriano, J., Mor, G., & Madsen, H. (2016). Hidden Markov Models for indirect classification of occupant behaviour. *Sustainable Cities and Society*, *27*, 83–98. https://doi.org/https://doi.org/10.1016/j.scs.2016.07.001

Linderman, G. C., & Steinerberger, S. (2019). Clustering with t-SNE, Provably. *SIAM Journal on Mathematics of Data Science*, *1*(2). https://doi.org/10.1137/18m1216134

Lomas, K. J., & Kane, T. (2013). Summertime temperatures and thermal comfort in UK homes. *Building Research and Information*, *41*(3), 259–280. https://doi.org/10.1080/09613218.2013.757886

Lu, Y., Wang, S., Yan, C., & Huang, Z. (2017). Robust optimal design of renewable energy system in nearly/net zero energy buildings under uncertainties. *Applied Energy*, *187*. https://doi.org/10.1016/j.apenergy.2016.11.042

Ma, Z., Cooper, P., Daly, D., & Ledo, L. (2012). Existing building retrofits: Methodology and state-of-the-art. *Energy and Buildings*, *55*(February 2018), 889–902. https://doi.org/10.1016/j.enbuild.2012.08.018

Mahdavi, A., & Tahmasebi, F. (2015). Predicting people's presence in buildings: An empirically based model performance analysis. *Energy and Buildings*, *86*. https://doi.org/10.1016/j.enbuild.2014.10.027

Mehregan, M., Naminezhad, A., Vakili, S., & Delpisheh, M. (2022). Building energy model validation and estimation using heating and cooling degree days (HDD–CDD) based on accurate base temperature. *Energy Science and Engineering*, *10*(9). https://doi.org/10.1002/ese3.1246

Melfi, R., Rosenblum, B., Nordman, B., & Christensen, K. (2011). Measuring building occupancy using existing network infrastructure. *2011 International Green Computing Conference and Workshops, IGCC 2011*. https://doi.org/10.1109/IGCC.2011.6008560

Micolier, A., Taillandier, F., Taillandier, P., & Bos, F. (2019). Li-BIM, an agent-based approach to simulate occupant-building interaction from the Building-Information Modelling. *Engineering Applications of Artificial Intelligence*. https://doi.org/10.1016/j.engappai.2019.03.008

Migliori, M., Najafi, H., Fabregas, A., & Nguyen, T. (2022). Neural Network-Based Building Energy Models for Adapting to Post-Occupancy Conditions: A Case Study for Florida. *Journal of Engineering for Sustainable Buildings and Cities*, *3*(4). https://doi.org/10.1115/1.4056393

Mitra, D., Steinmetz, N., Chu, Y., & Cetin, K. S. (2020). Typical occupancy profiles and behaviors in residential buildings in the United States. *Energy and Buildings*, *210*, 109713. https://doi.org/10.1016/j.enbuild.2019.109713

Mosteiro-Romero, M., Hischier, I., Fonseca, J. A., & Schlueter, A. (2020). A novel population-based occupancy modeling approach for district-scale simulations compared to standard-based methods. *Building and Environment*, *181*, 107084. https://doi.org/10.1016/j.buildenv.2020.107084

Nejadshamsi, S., Eicker, U., Wang, C., & Bentahar, J. (2023). Data sources and approaches for building occupancy profiles at the urban scale – A review. *Building and Environment*, *238*, 110375. https://doi.org/https://doi.org/10.1016/j.buildenv.2023.110375

Nguyen, V. (2019). Bayesian optimization for accelerating hyper-parameter tuning. *Proceedings - IEEE 2nd International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2019*. https://doi.org/10.1109/AIKE.2019.00060

Noh, S. H. (2021). Analysis of gradient vanishing of RNNs and performance comparison. *Information (Switzerland)*, *12*(11). https://doi.org/10.3390/info12110442

Öztürk, M. Z., Çetinkaya, G., & Aydın, S. (2017). Climate Types of Turkey According to Köppen-Geiger Climate Classification. *Istanbul University Journal of Geography*, *35*.

Pacheco, R., Ordóñez, J., & Martínez, G. (2012). Energy efficient design of building: A review. In *Renewable and Sustainable Energy Reviews* (Vol. 16, Issue 6). https://doi.org/10.1016/j.rser.2012.03.045

Padilla-López, J. R., Chaaraoui, A. A., & Flórez-Revuelta, F. (2015). Visual privacy protection methods: A survey. *Expert Systems with Applications*, *42*(9), 4177–4195. https://doi.org/https://doi.org/10.1016/j.eswa.2015.01.041

Page, J., Robinson, D., Morel, N., & Scartezzini, J. L. (2008). A generalised stochastic model for the simulation of occupant presence. *Energy and Buildings*, *40*(2), 83–98. https://doi.org/10.1016/j.enbuild.2007.01.018

Pan, Y., Zhu, M., Lv, Y., Yang, Y., Liang, Y., Yin, R., Yang, Y., Jia, X., Wang, X., Zeng, F., Huang, S., Hou, D., Xu, L., Yin, R., & Yuan, X. (2023). Building energy simulation and its application for building performance optimization: A review of methods, tools, and case studies. *Advances in Applied Energy*, *10*, 100135. https://doi.org/https://doi.org/10.1016/j.adapen.2023.100135

Panchabikesan, K., Haghighat, F., & Mankibi, M. El. (2021a). Data driven occupancy information for energy simulation and energy use assessment in residential buildings. *Energy*, *218*. https://doi.org/10.1016/j.energy.2020.119539

Panchabikesan, K., Haghighat, F., & Mankibi, M. El. (2021b). Data driven occupancy information for energy simulation and energy use assessment in residential buildings. *Energy*, *218*. https://doi.org/10.1016/j.energy.2020.119539

Pang, Z., Xu, P., O'Neill, Z., Gu, J., Qiu, S., Lu, X., & Li, X. (2018). Application of mobile positioning occupancy data for building energy simulation: An engineering case study. *Building and Environment*, *141*. https://doi.org/10.1016/j.buildenv.2018.05.030

Park, J. Y., Dougherty, T., & Intelligent, Z. N. (2018). A bluetooth based occupancy detection for buildings. *ASHRAE and IBPSA-USA Building Simulation Conference*.

Parker, J., Hardy, A., Glew, D., & Gorse, C. (2017). A methodology for creating building energy model occupancy schedules using personal location metadata. *Energy and Buildings*, *150*. https://doi.org/10.1016/j.enbuild.2017.06.014

Petersen, S., Pedersen, T. H., Nielsen, K. U., & Knudsen, M. D. (2016). Establishing an image-based ground truth for validation of sensor data-based room occupancy detection. *Energy and Buildings*, *130*. https://doi.org/10.1016/j.enbuild.2016.09.009

Petrovic, S. (2006). A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters. *11th Nordic Workshop on Secure IT-Systems*.

Philip, S., Tran, T., & Tanjuatco, L. (2011a). eppy: scripting language for E+. *EnergyPlus (Version 0.46)[Software-GNU AFFERO GENERAL PUBLIC LICENSE] Avaliable From< Https://Pypi. Python. Org/Pypi/Eppy/0.4*, *6*.

Philip, S., Tran, T., & Tanjuatco, L. (2011b). eppy: scripting language for E+. *EnergyPlus (Version 0.46)[Software-GNU AFFERO GENERAL PUBLIC LICENSE] Avaliable From< Https://Pypi. Python. Org/Pypi/Eppy/0.4*, *6*.

Picaut, J., Fortin, N., Bocher, E., Petit, G., Aumond, P., & Guillaume, G. (2019). An open-science crowdsourcing approach for producing community noise maps using smartphones. *Building and Environment*, *148*. https://doi.org/10.1016/j.buildenv.2018.10.049

Picco, M., Lollini, R., & Marengo, M. (2014). Towards energy performance evaluation in early stage building design: A simplification methodology for commercial building models. *Energy and Buildings*, *76*, 497–505.

Potdar, K., S., T., & D., C. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, *175*(4). https://doi.org/10.5120/ijca2017915495

Quan, S. J., Li, Q., Augenbroe, G., Brown, J., & Yang, P. P. J. (2015). Urban data and building energy modeling: A GIS-based urban building energy modeling system using the urban-EPC engine. *Lecture Notes in Geoinformation and Cartography*, *213*. https://doi.org/10.1007/978-3-319-18368-8_24

Raykov, Y. P., Ozer, E., Dasika, G., Boukouvalas, A., & Little, M. A. (2016). Predicting room occupancy with a single Passive infrared (PIR) sensor through behavior extraction. *UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. https://doi.org/10.1145/2971648.2971746

Reinhart, C. F., & Cerezo Davila, C. (2016). Urban building energy modeling - A review of a nascent field. *Building and Environment*, *97*, 196–202. https://doi.org/10.1016/j.buildenv.2015.12.001

Ren, X., Yan, D., & Hong, T. (2015). Data mining of space heating system performance in affordable housing. *Building and Environment*, *89*. https://doi.org/10.1016/j.buildenv.2015.02.009

Richardson, I., Thomson, M., & Infield, D. (2008). A high-resolution domestic building occupancy model for energy demand simulations. *Energy and Buildings*, *40*(8), 1560–1566. https://doi.org/https://doi.org/10.1016/j.enbuild.2008.02.006

Richardson, I., Thomson, M., Infield, D., & Delahunty, A. (2009). Domestic lighting: A high-resolution energy demand model. *Energy and Buildings*, *41*(7). https://doi.org/10.1016/j.enbuild.2009.02.010

Robinson, D., Campbell, N., Gaiser, W., Kabel, K., Le-Mouel, A., Morel, N., Page, J., Stankovic, S., & Stone, A. (2007). SUNtool - A new modelling paradigm for simulating and optimising urban sustainability. *Solar Energy*, *81*(9). https://doi.org/10.1016/j.solener.2007.06.002

Saitta, S., Raphael, B., & Smith, I. F. C. (2008). A comprehensive validity index for clustering. *Intelligent Data Analysis*, *12*(6). https://doi.org/10.3233/ida-2008-12602

Santin, O. G. (2011). Behavioural patterns and user profiles related to energy consumption for heating. *Energy and Buildings*, *43*(10). https://doi.org/10.1016/j.enbuild.2011.06.024

Schweiker, M., Carlucci, S., Andersen, R. K., Dong, B., & O'Brien, W. (2017). Occupancy and occupants' actions. In *Exploring Occupant Behavior in Buildings: Methods and Challenges* (pp. 7–38). Springer International Publishing. https://doi.org/10.1007/978-3-319-61464-9_2

Shao, W., Salim, F. D., Nguyen, T., & Youssef, M. (2017). Who Opened the Room? Device-Free Person Identification Using Bluetooth Signals in Door Access. *Proceedings - 2017 IEEE International Conference on Internet of Things, IEEE Green Computing and Communications, IEEE Cyber, Physical and Social Computing, IEEE Smart Data, IThings-GreenCom-CPSCom-SmartData 2017*, *2018-January*. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.16

Sheikh Khan, D., Kolarik, J., Anker Hviid, C., & Weitzmann, P. (2021). Method for long-term mapping of occupancy patterns in open-plan and single office spaces by using passive-infrared (PIR) sensors mounted below desks. *Energy and Buildings*, *230*. https://doi.org/10.1016/j.enbuild.2020.110534

Shetty, P., & Singh, S. (2021). Hierarchical Clustering: A Survey. *International Journal of Applied Research*, *7*(4). https://doi.org/10.22271/allresearch.2021.v7.i4c.8484

SIA. (2006). *Conditions d'utilisation standard pour l'énergie et les installations du bâtiment, swiss society engineers and architects*.

Smil, V. (2017). Energy and Civilization: A History. In *The MIT Press*. The MIT Press.

Snodgrass, S., & Ontañón, S. (2014). Experiments in map generation using Markov chains. *Proceedings of the 9th International Conference on Foundations of Digital Games*.

Sood, D., Alhindawi, I., Ali, U., McGrath, J. A., Byrne, M. A., Finn, D., & O'Donnell, J. (2023). Simulation-based evaluation of occupancy on energy consumption of multi-scale residential building archetypes. *Journal of Building Engineering*, *75*. https://doi.org/10.1016/j.jobe.2023.106872

Sudarno, P. W., Ashari, A., & Riasetiawan, M. (2022). RAINFALL ANALYSIS IN SEMARANG CITY USING K-MEANS AND AGGLOMERATIVE HIERARCHICAL CLUSTERING METHODS. *ICIC Express Letters, Part B: Applications*, *13*(11). https://doi.org/10.24507/icicelb.13.11.1199

Tamer, T., Gursel Dino, I., & Meral Akgul, C. (2022). Data-driven, long-term prediction of building performance under climate change: Building energy demand and BIPV energy generation analysis across Turkey. *Renewable and Sustainable Energy Reviews*, *162*, 112396. https://doi.org/10.1016/j.rser.2022.112396

Tardioli, G., Kerrigan, R., Oates, M., O'Donnell, J., & Finn, D. P. (2018). Identification of representative buildings and building groups in urban datasets using a novel pre-processing, classification, clustering and predictive modelling approach. *Building and Environment*, *140*, 90–106. https://doi.org/https://doi.org/10.1016/j.buildenv.2018.05.035

Thomas, D., Miller, C., Kämpf, J., & Schlueter, A. (2014). Multiscale Co-simulation of EnergyPlus and CitySim models derived from a Building Information Model. *Bausim 2014: Fifth German-Austrian IBPSA Conference.*

Tomastik, R., Lin, Y., & Banaszuk, A. (2008). Video-based estimation of building occupancy during emergency egress. *Proceedings of the American Control Conference*. https://doi.org/10.1109/ACC.2008.4586606

Torriti, J. (2014). A review of time use models of residential electricity demand. In *Renewable and Sustainable Energy Reviews* (Vol. 37). https://doi.org/10.1016/j.rser.2014.05.034

Torriti, J. (2017). Understanding the timing of energy demand through time use data: Time of the day dependence of social practices. *Energy Research and Social Science*, *25*. https://doi.org/10.1016/j.erss.2016.12.004

TUIK. (2022). *Turkish Statistical institute, Data Portal for Statistics*.

Tuniki, H. P., Jurelionis, A., & Fokaides, P. (2021). A review on the approaches in analysing energy-related occupant behaviour research. *Journal of Building Engineering*, *40*, 102630. https://doi.org/https://doi.org/10.1016/j.jobe.2021.102630

Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*.

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, *53*(8). https://doi.org/10.1007/s10462-020-09838-1 van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin and Review*, *25*(1). https://doi.org/10.3758/s13423-016-1015-8

Virote, J., & Neves-silva, R. (2012). Stochastic models for building energy prediction based on occupant behavior assessment. *Energy & Buildings*, *53*, 183–193. https://doi.org/10.1016/j.enbuild.2012.06.001

Walker, G. (2014). The dynamics of energy demand: Change, rhythm and synchronicity. *Energy Research and Social Science*, *1*. https://doi.org/10.1016/j.erss.2014.03.012

Wang, C., Ferrando, M., Causone, F., Jin, X., Zhou, X., & Shi, X. (2022). Data acquisition for urban building energy modeling: A review. *Building and Environment*, *217*(January), 109056. https://doi.org/10.1016/j.buildenv.2022.109056

Wang, M., Wang, X., Zhang, G., & Li, C. (2015). Occupancy detection based on Spiking Neural Networks for green building automation systems. *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)*, *2015-March*(March). https://doi.org/10.1109/WCICA.2014.7053149

Wang, Y., & Shao, L. (2017). Understanding occupancy pattern and improving building energy efficiency through Wi-Fi based indoor positioning. *Building and Environment*, *114*. https://doi.org/10.1016/j.buildenv.2016.12.015

Wei, S., Jones, R., & De Wilde, P. (2014). Driving factors for occupant-controlled space heating in residential buildings. In *Energy and Buildings* (Vol. 70). https://doi.org/10.1016/j.enbuild.2013.11.001

Widén, J., Nilsson, A. M., & Wäckelgård, E. (2009). A combined Markov-chain and bottom-up approach to modelling of domestic lighting demand. *Energy and Buildings*, *41*(10). https://doi.org/10.1016/j.enbuild.2009.05.002

Wilke, U., Haldi, F., Scartezzini, J. L., & Robinson, D. (2013). A bottom-up stochastic model to predict building occupants' time-dependent activities. *Building and Environment*, *60*. https://doi.org/10.1016/j.buildenv.2012.10.021

Winkelmann, D. B. C., Pedersen, C. O., Lawrie, L. K., & C., F. (2000). EnergyPlus: Energy Simulation Program. *ASHRAE Journal*, *42*, 49–56.

Wolf, S., Cali, D., Alonso, M. J., Li, R., Andersen, R. K., Krogstie, J., & Madsen, H. (2019). Room-level occupancy simulation model for private households. *Journal of Physics: Conference Series*, *1343*(1). https://doi.org/10.1088/1742-6596/1343/1/012126

Xu, B., Fu, L., & Di, H. (2009). Field investigation on consumer behavior and hydraulic performance of a district heating system in Tianjin, China. *Building and Environment*, *44*(2). https://doi.org/10.1016/j.buildenv.2008.03.002

Yan, D., Hong, T., Dong, B., Mahdavi, A., D'Oca, S., Gaetani, I., & Feng, X. (2017). IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings. *Energy and Buildings*, *156*, 258–270. https://doi.org/https://doi.org/10.1016/j.enbuild.2017.09.084

Yan, D., O'Brien, W., Hong, T., Feng, X., Burak Gunay, H., Tahmasebi, F., & Mahdavi, A. (2015). Occupant behavior modeling for building performance simulation: Current state and future challenges. *Energy and Buildings*, *107*, 264–278. https://doi.org/https://doi.org/10.1016/j.enbuild.2015.08.032

Yang, B., Haghighat, F., Fung, B. C. M., & Panchabikesan, K. (2021). Season-Based Occupancy Prediction in Residential Buildings Using Machine Learning Models. *E-Prime - Advances in Electrical Engineering, Electronics and Energy*, *1*. https://doi.org/10.1016/j.prime.2021.100003

Yang, J., Ning, C., Deb, C., Zhang, F., Cheong, D., Lee, S. E., Sekhar, C., & Tham, K. W. (2017). k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy and Buildings*, *146*. https://doi.org/10.1016/j.enbuild.2017.03.071

Yang, J., Santamouris, M., & Lee, S. E. (2016). Review of occupancy sensing systems and occupancy modeling methodologies for the application in institutional buildings. *Energy and Buildings*. https://doi.org/10.1016/j.enbuild.2015.12.019

Yang, X., Hu, M., Zhang, C., & Steubing, B. (2022). Key strategies for decarbonizing the residential building stock: Results from a spatiotemporal model for Leiden, the Netherlands. *Resources, Conservation and Recycling*, *184*, 106388. https://doi.org/https://doi.org/10.1016/j.resconrec.2022.106388

Yao, J. (2018). Modelling and simulating occupant behaviour on air conditioning in residential buildings. *Energy and Buildings*, *175*. https://doi.org/10.1016/j.enbuild.2018.07.013

Yoshino, H., Hong, T., Nord, N., & Zhang, X. (2017). IEA EBC annex 53: Total energy use in buildings—Analysis and evaluation methods. *Energy and Buildings*, *152*(July), 124–136. https://doi.org/10.1016/j.enbuild.2017.07.038

Yu, Z., Fung, B. C. M., Haghighat, F., Yoshino, H., & Morofsky, E. (2011). A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy and Buildings*, *43*(6). https://doi.org/10.1016/j.enbuild.2011.02.002

Yu, Z. (Jerry), Haghighat, F., & Fung, B. C. M. (2016). Advances and challenges in building engineering and data mining applications for energy-efficient communities. *Sustainable Cities and Society*, *25*, 33–38. https://doi.org/https://doi.org/10.1016/j.scs.2015.12.001

Zhang, S., Guo, Q., Smyth, R., & Yao, Y. (2022). Extreme temperatures and residential electricity consumption: Evidence from Chinese households.

Energy Economics, *107*, 105890. https://doi.org/https://doi.org/10.1016/j.eneco.2022.105890

Zhao, J., Lasternas, B., Lam, K. P., Yun, R., & Loftness, V. (2014). Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining. *Energy and Buildings*, *82*. https://doi.org/10.1016/j.enbuild.2014.07.033

Zhou, X., Lu, Y., Hu, S., Yang, Z., & Yan, D. (2023). New perspectives on temporal changes in occupancy characteristics of residential buildings. *Journal of Building Engineering*, *64*, 105590. https://doi.org/https://doi.org/10.1016/j.jobe.2022.105590

Zhou, X., Yan, D., Hong, T., & Ren, X. (2015). Data analysis and stochastic modeling of lighting energy use in large office buildings in China. *Energy and Buildings*, *86*. https://doi.org/10.1016/j.enbuild.2014.09.071

Zhuang, C., Choudhary, R., & Mavrogianni, A. (2022). Probabilistic occupancy forecasting for risk-aware optimal ventilation through autoencoder Bayesian deep neural networks. *Building and Environment*, *219*. https://doi.org/10.1016/j.buildenv.2022.109207

Zou, J., Zhao, Q., Yang, W., & Wang, F. (2017). Occupancy detection in the office by analyzing surveillance videos and its application to building energy conservation. *Energy and Buildings*. https://doi.org/10.1016/j.enbuild.2017.07.064

APPENDICES

A. Tuning Scores for ML with Feature Engineering

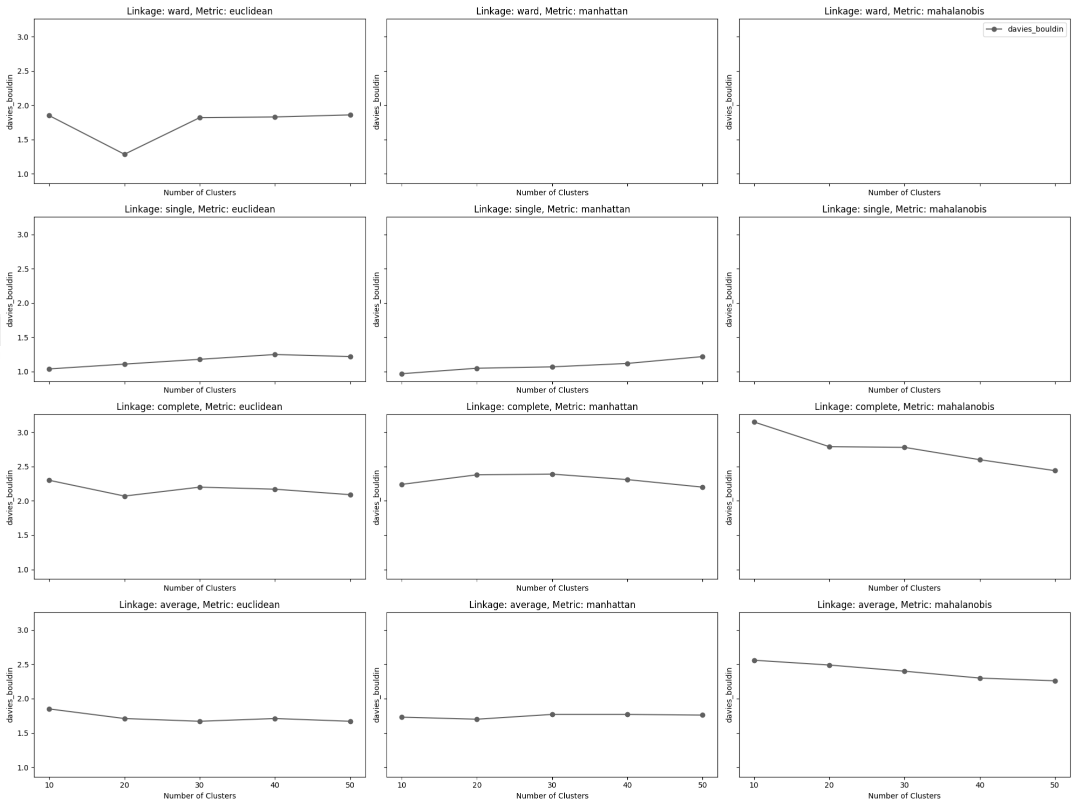

Figure A. 1 – Davies Bouldin score over hyper-parameter space, initial phase

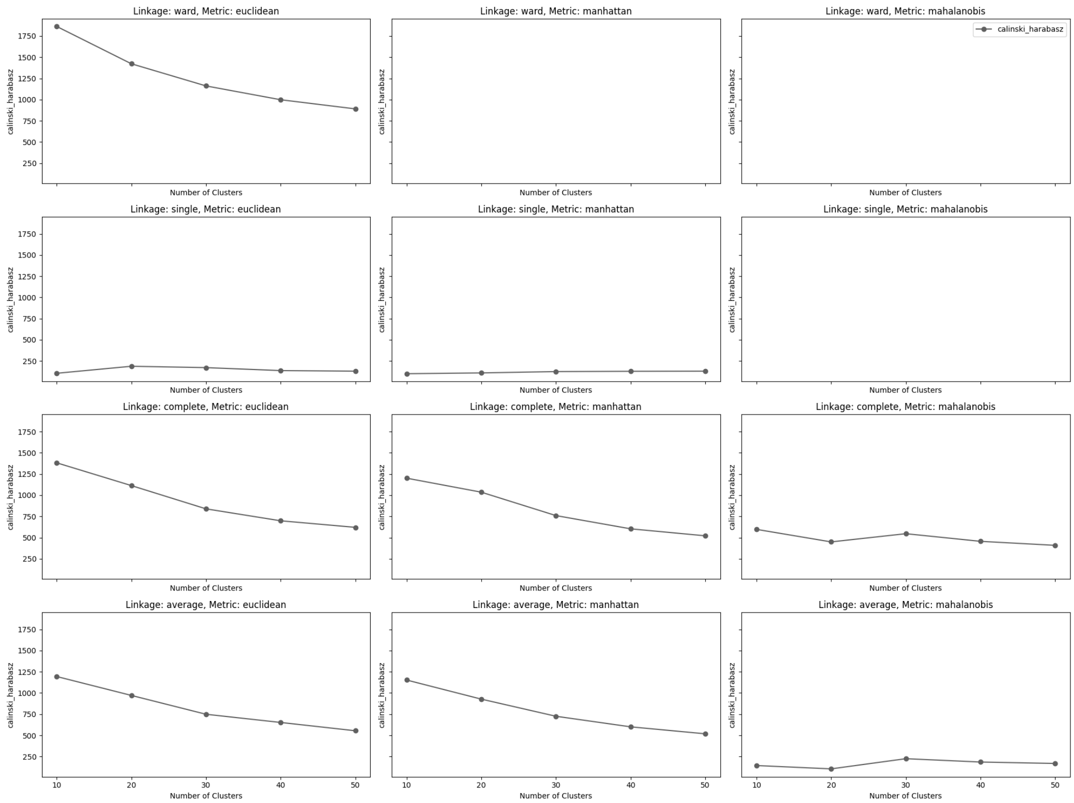

Figure A. 2 – Calinski Harabasz score over hyper-parameter space, initial phase

B. End-to-End Deep Learning without Feature Engineering, LSTM Model Architectures

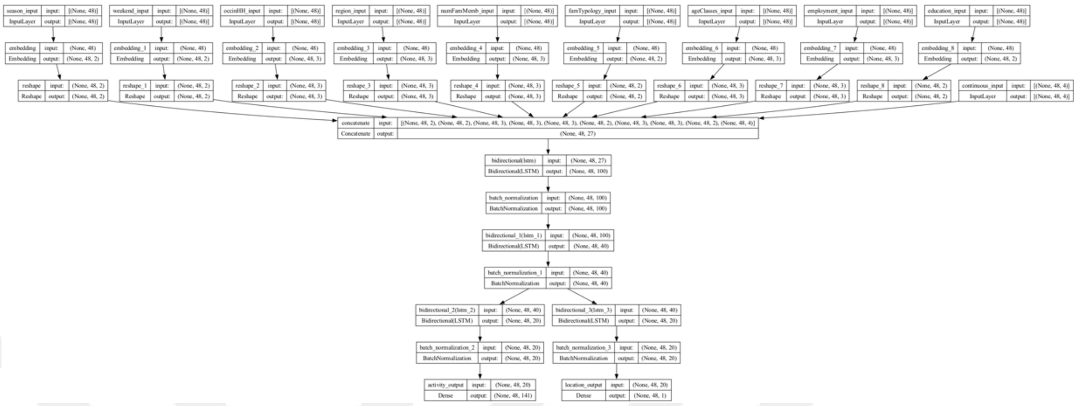

Figure B. 1 – LSTM Model Architecture, Before Tuning, two hidden layers

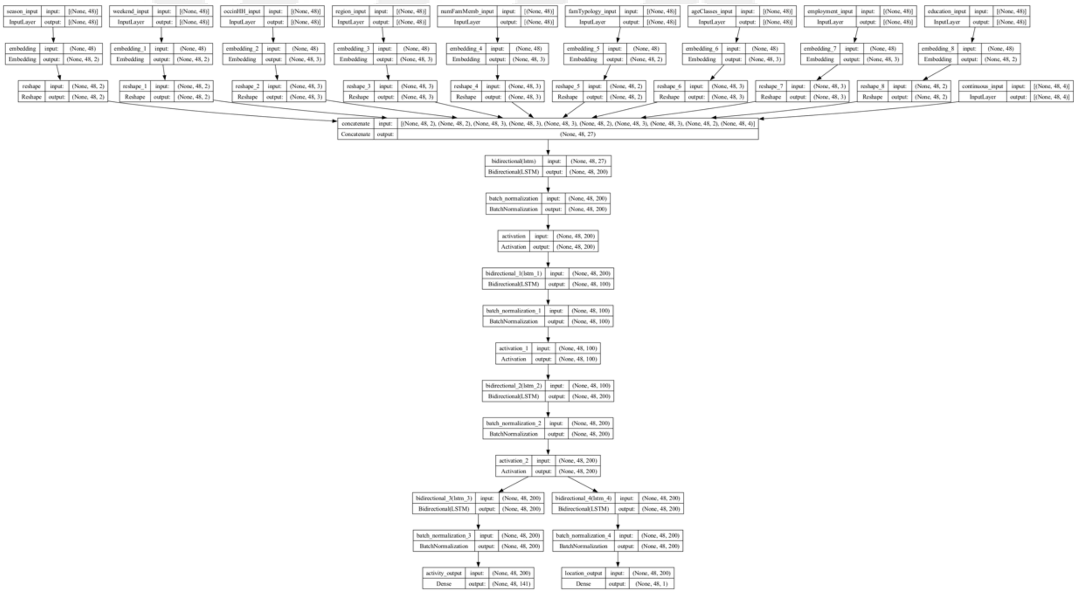

Figure B. 2 – LSTM Model Architecture, After Tuning, five hidden layers

C. Chi-Square test for Occupant's Presence Schedules: A Statistical Differentiation Test for ML with Feature Engineering Results

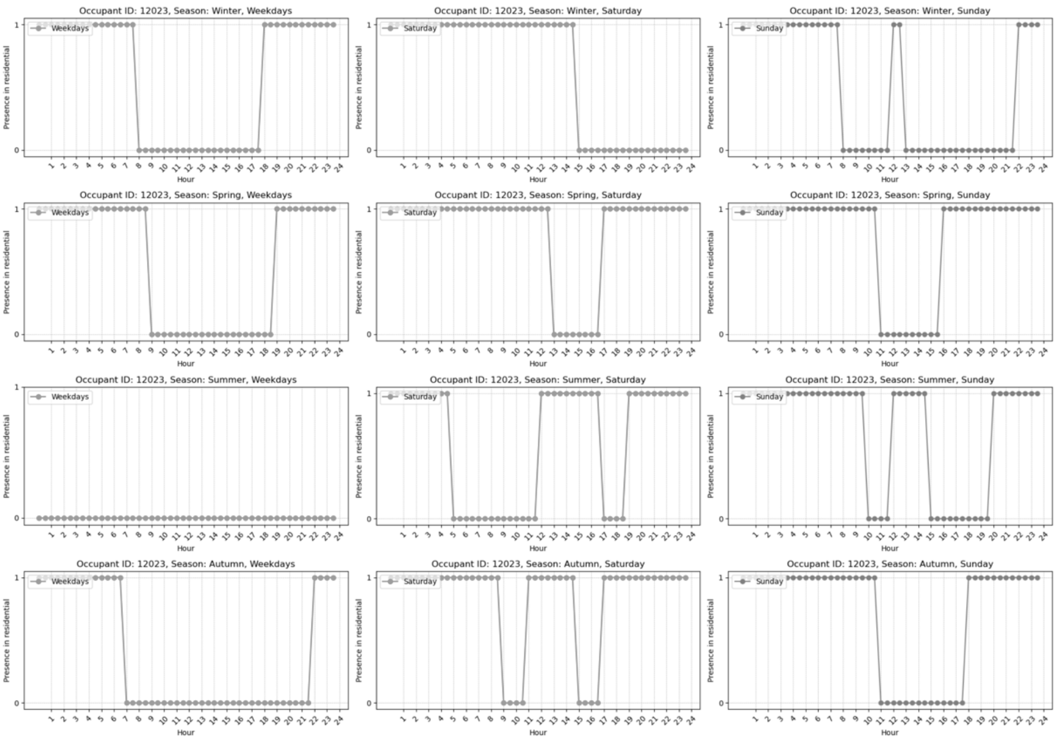

Figure C. 1 – Chi-square results (i.e. p-value) of presence schedule of an occupant: for Weekdays is 5.674, for Saturday is 0.0144, for Sunday is 0.00367

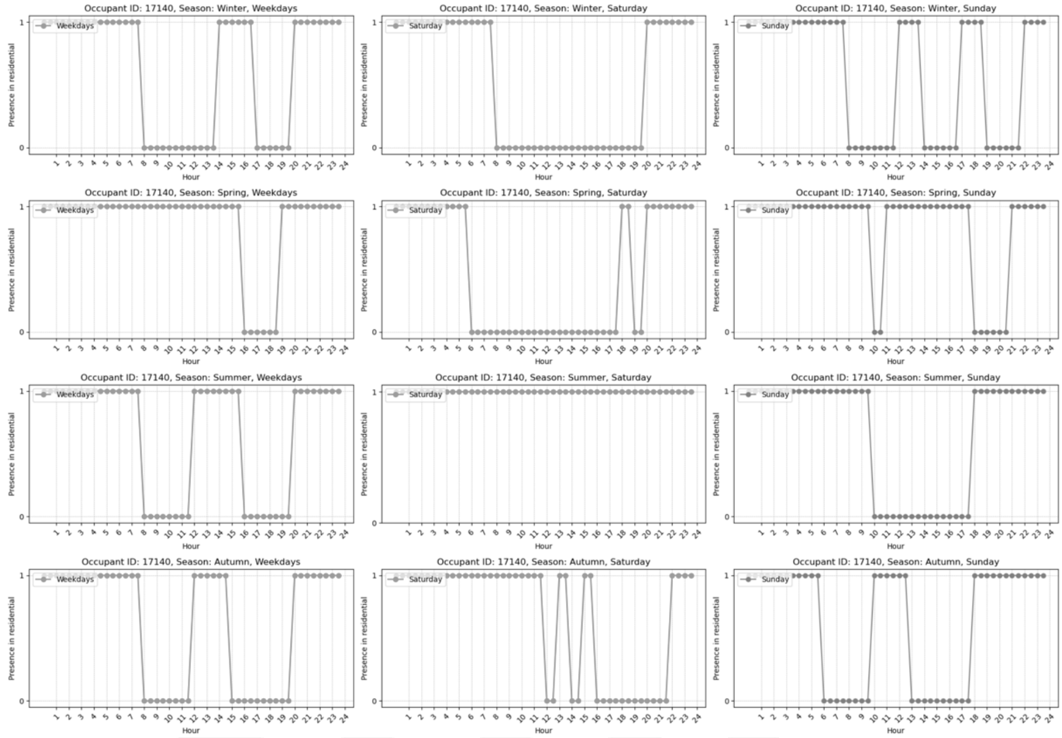

Figure C. 2 – Chi-square results (i.e. p-value) of presence schedule of an occupant: for Weekdays is 0.02050, for Saturday is 2.01354, for Sunday is 0.04793

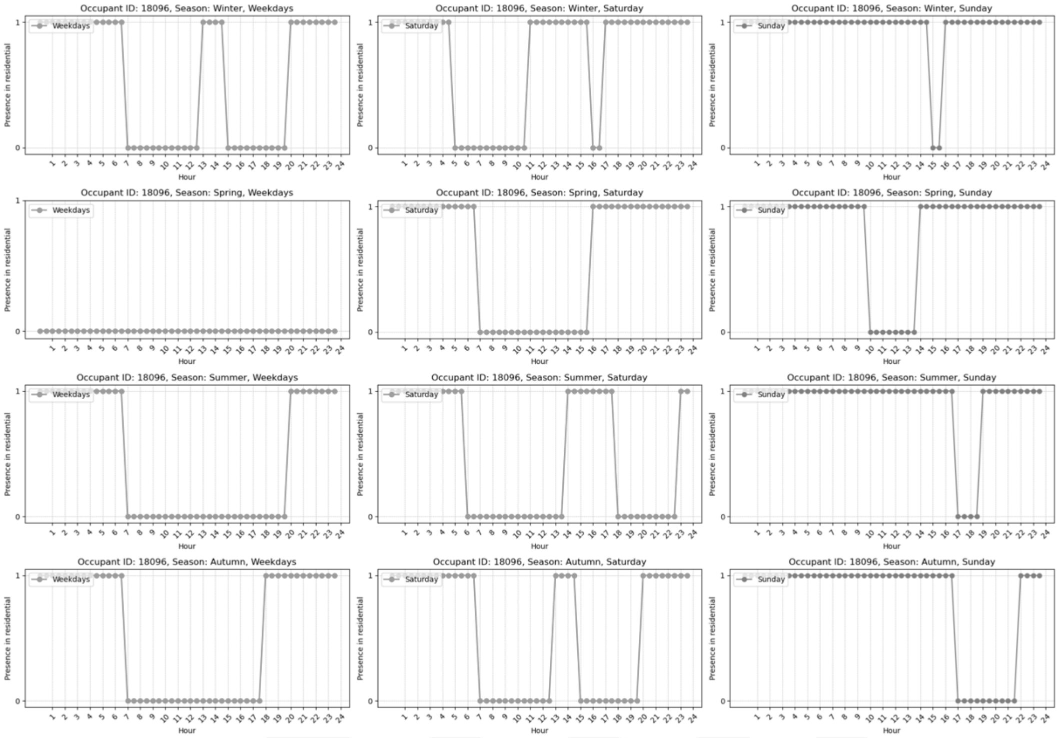

Figure C. 3 – Chi-square results (i.e. p-value) of presence schedule of an occupant: for Weekdays is 6.3078, for Saturday is 0.0765, for Sunday is 0.05457

D. Chi-Square test for Occupant's Presence Schedules: A Statistical Differentiation Test for End to End without Feature Engineering Results

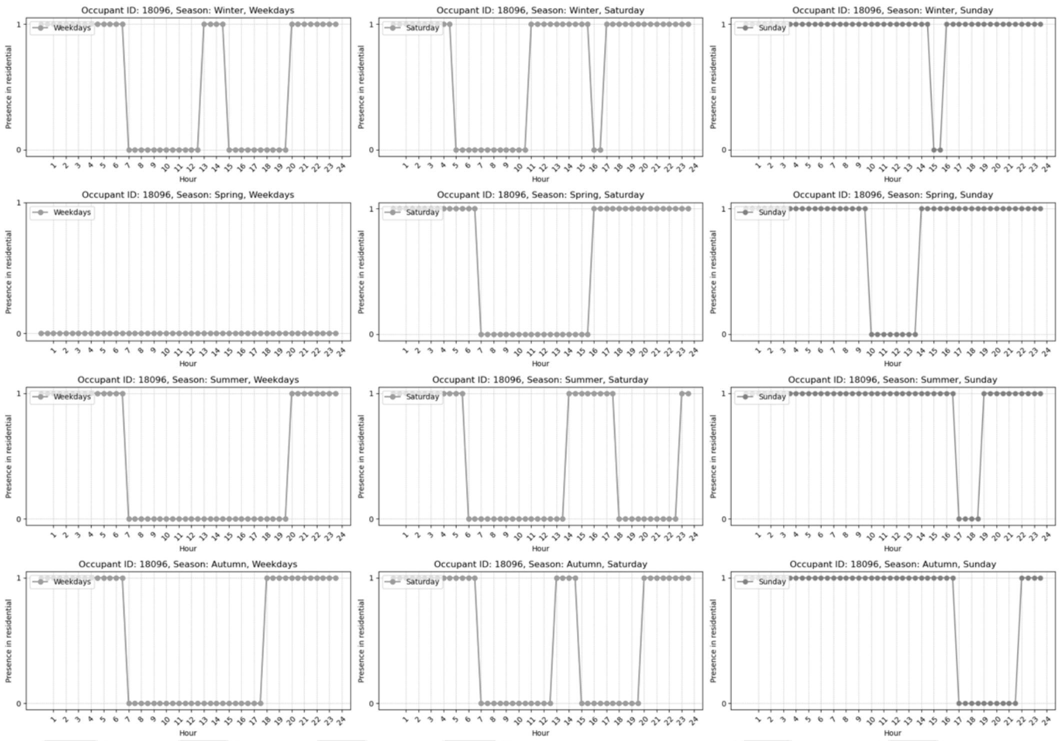

Figure D. 1 – Chi-square results (i.e. p-value) of presence schedule of an occupant: for Weekdays is 0.0269, for Saturday is 0.07004, for Sunday is 0.01966

E. LSTM Multi-categorical Performance Analysis For Activity Schedule Category Generation In Terms of Precision, Recall And F1-score

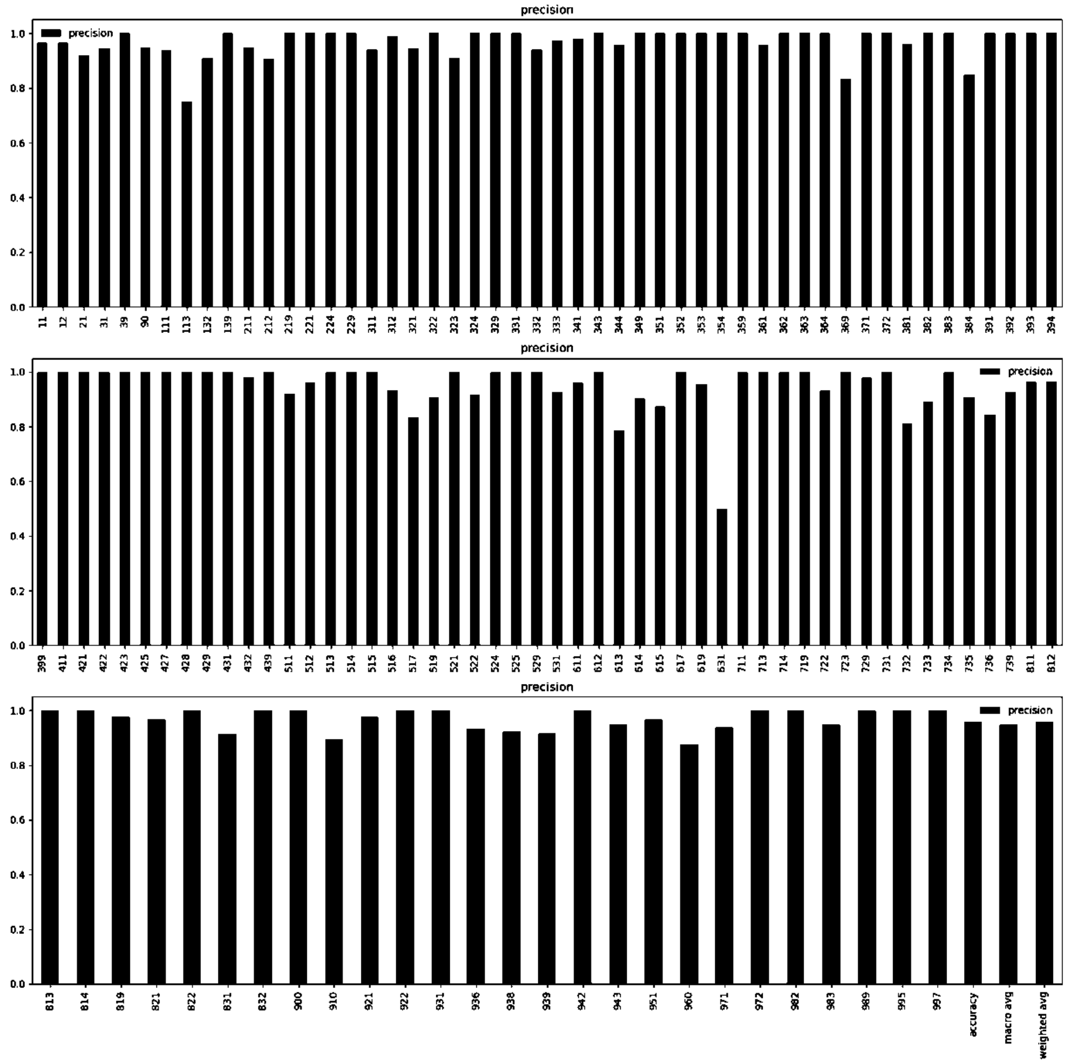

Figure E. 1 – Precision score results of "Occupant_Activity" column

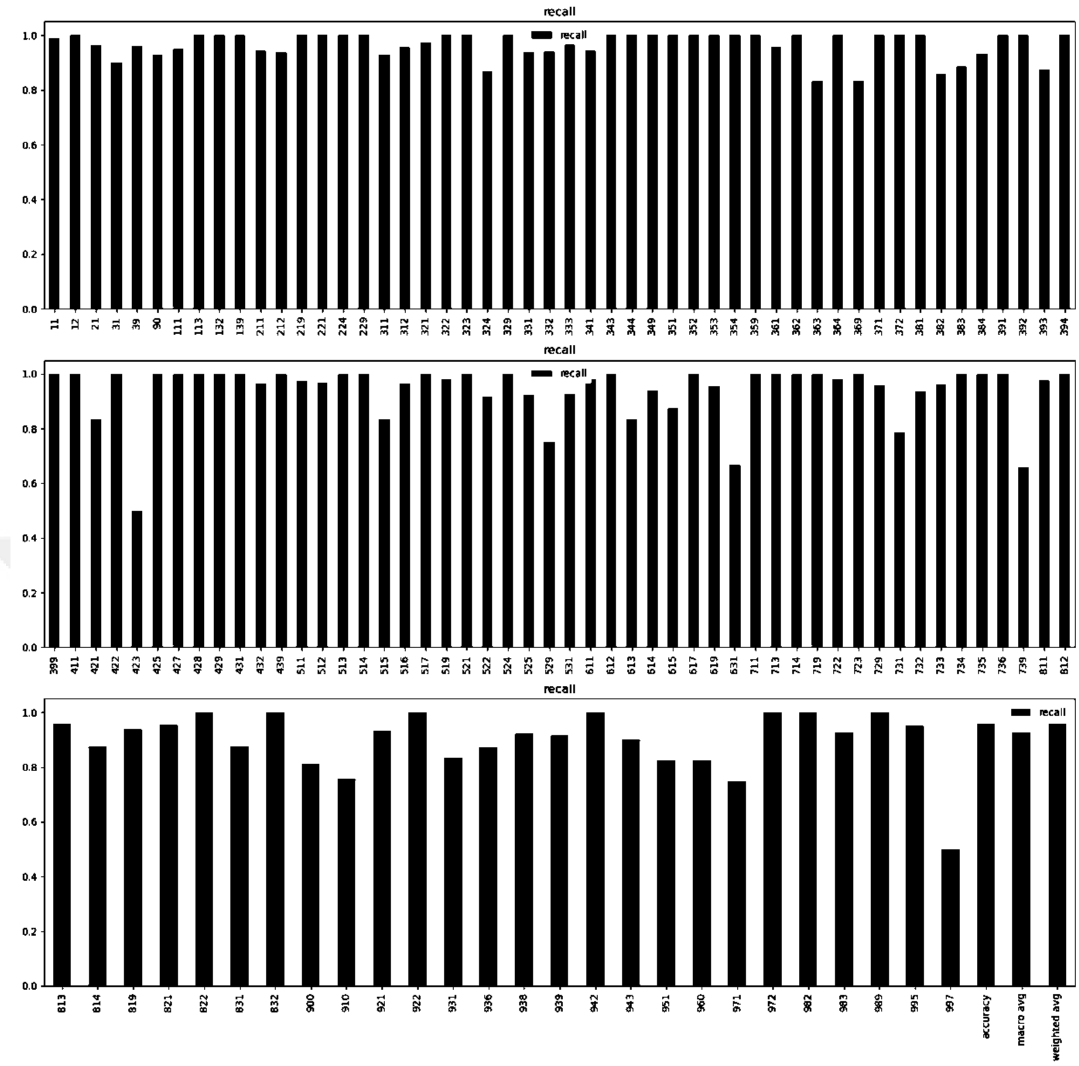

Figure E. 2 – Recall score results of “Occupant_Activity” column

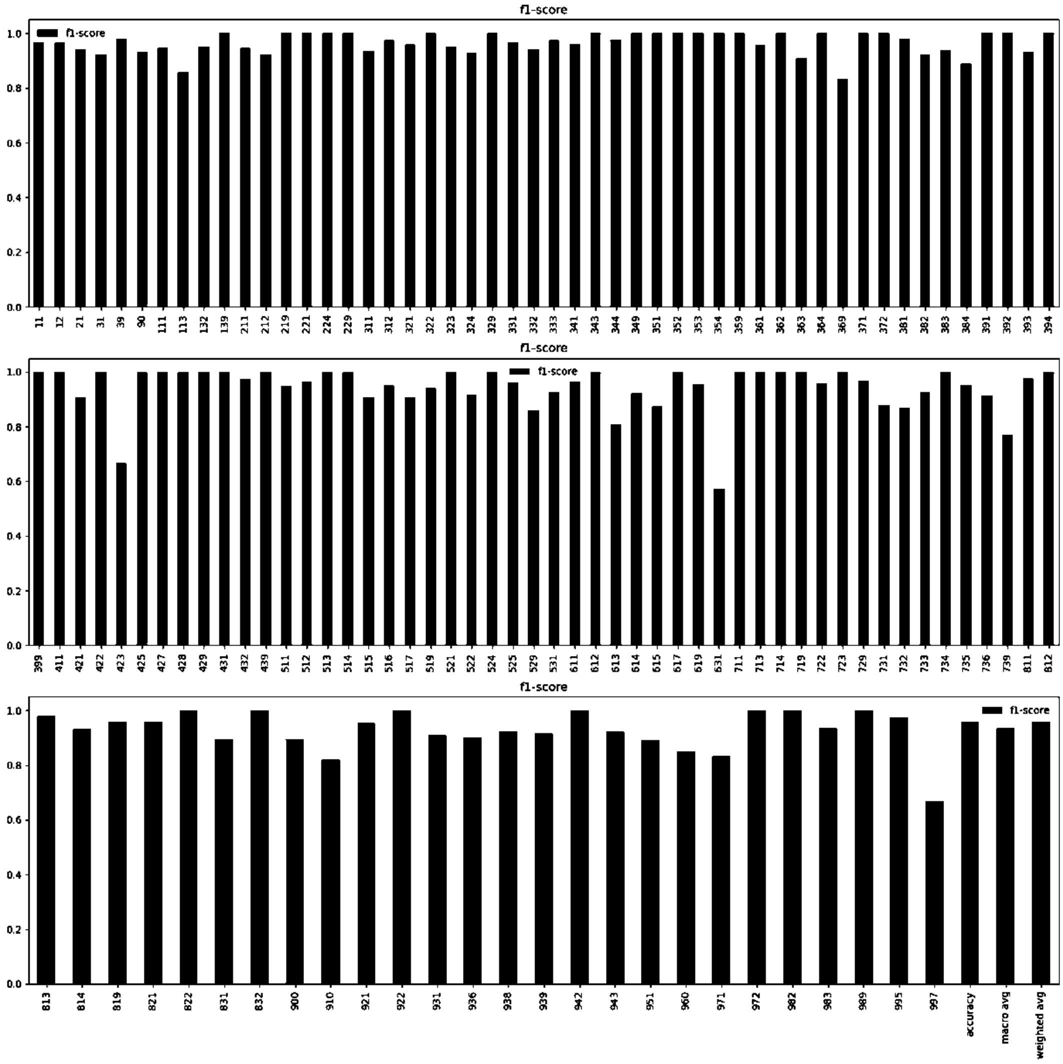

Figure E. 3 – F1-score results of “Occupant_Activity” column

CURRICULUM VITAE

Surname, Name: İşeri, Orçun Koral

EDUCATION

Degree	Institution	Year of Graduation
MS	Yasar University, Architecture	2018
BS	Yasar University, Architecture	2016
High School	Lycée Saint-Joseph, Izmir	2011

FOREIGN LANGUAGES

Advanced English, Fluent French, Beginner German, and Italian

PUBLICATIONS

1. Akin, S., Iseri, O. K., Dino, I. G., & Erdogan, B. (2022). Exploratory and Multi-Objective Decision-Making Methods for Retrofit Planning Processes. International Journal of Digital Innovation in the Built Environment (IJDIBE), 11(2), 1–16.
2. AKIN, Ş., İŞERİ, O. K., ERDOĞAN, B., & DİNO, İ. G. (2020). An exploratory multi-objective retrofit decision-making process. ATI.
3. Akın, Ş., İşeri, O. K., Akgül, Ç. M., Erdoğan, B., & Dino, İ. G. (2020). Climate change impact on multi-objective optimization: a case study on educational buildings. Proceedings of the 11th Annual Symposium on Simulation for Architecture and Urban Design, 1–8.
4. Canlı, İ., İşeri, O. K., & Dino, İ. G. (2021). Comparision of Different Data-Driven Models on Prediction of Useful Daylight Illuminance (UDI). EasyChair.
5. Dino, I. G., Kalfaoglu, E., Iseri, O. K., Erdogan, B., Kalkan, S., & Alatan, A. A. (2022). Vision-based estimation of the number of occupants using video cameras. Advanced Engineering Informatics, 53, 101662.
6. Dino, I. G., Kalfaoglu, E., Sarı, A. E., ANin, S., Iseri, O. K., Alatan, A. A., KalNan, S., & Erdogan, B. (2019). Automated building energy modeling

for existing buildings using computer vision. 36th CIB W78 2019 Conference, 851–862.

7. Dino, I. G., Sari, A. E., Iseri, O. K., Akin, S., Kalfaoglu, E., Erdogan, B., Kalkan, S., & Alatan, A. A. (2020). Image-based construction of building energy models using computer vision. Automation in Construction, 116, 103231.
8. Dino, I., Kalfaoglu, E., Sari, A. E., Akin, S., Iseri, O., Alatan, A., Kalkan, S., & Erdogan, B. (2019). Video content analysis-based detection of occupant presence for building energy modelling. Proceedings of ICT in Design, Construction and Management in Architecture, Engineering, Construction and Operations (AECO), Newcastle, UK.
9. Duran, A., Işeri, O. K., Akgül, Ç., Kalkan, S., & Gürsel Dino, İ. (2022). Compiling open datasets to improve urban building energy models with occupancy and layout data.
10. Gursel Dino, I., Kalfaoglu, E., Iseri, O. K., Erdogan, B., Kalkan, S., & Alatan, A. A. (2022). Vision-based estimation of the number of occupants using video cameras.
11. Gürsel Dino, İ., Sarı, A. E., Akın, Ş., Işeri, O. K., Alatan, A. A., & Kalkan, S. (2019). Video Content Analysis-Based Detection of Occupant Presence for Building Energy Modelling.
12. Gürsel Dino, I., Sarı, A. E., Kalfaoğlu, E., Akın, Ş., Işeri, O. K., Kalkan, S., Alatan, A. A., & Erdoğan, B. (2019). Automated building energy modeling for existing buildings using computer vision.
13. Halaçlı, E. G., Canlı, İ., İşeri, O. K., Yavuz, F., Akgül, Ç. M., Kalkan, S., & Dino, I. G. (2023). A Novel Graph Neural Network for Zone-Level Urban-Scale Building Energy Use Estimation. Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 169–176.

14. İşeri, O. K., Akin, S., & Dino, I. G. (2020). Energy Demand Prediction For Residential Buildings At Different Climate Conditions Based On Different Data-driven Models. Bausim 2020, TUGraz, 647–654.
15. İşeri, O. K., & Dino, İ. G. (2020). An Algorithm for Efficient Urban Building Energy Modeling and Simulation. SIMAUD - AUSTRIA.
16. İşeri, O. K., & Dino, İ. G. (2021). Building Archetype Characterization Using K-Means Clustering in Urban Building Energy Models. International Conference on Computer-Aided Architectural Design Futures, 222–236.
17. Iseri, O. K., & Dursun, O. (2020). The Comparative Study on the Influence of Early Architectural Design Decisions on Energy Demand: A Case Study in Turkey. Proceedings of ATI, 280–291.
18. İşeri, O. K., & Dursun, O. (2022). The Impacts of Early Architectural Design Decisions on Building Performance. International Journal of Digital Innovation in the Built Environment (IJDIBE), 11(2), 1–21.
19. İşeri, O. K., & Gürsel Dino, İ. (2020). Energy Demand Prediction For Residential Buildings At Different Climate Conditions Based On Different Data-driven Techniques.
20. Kalfaoglu, E., Dino, I. G., Iseri, O. K., Akin, S., Sari, A. E., Erdogan, B., Kalkan, S., & Alatan, A. (2020). Vision-Based Lighting State Detection and Curtain Openness Ratio Prediction. SIMAUD.
21. Uyduran, H. G., & İşeri, O. K. (2017). Net Present Value and UDI optimization with HypE Genetic Algortihm. MSTAS.
22. Uyduran, H. G., İşeri, O. K., Üstünes, Y., & Dursun, O. (2016). Optimizing wall insulation material parameters in renovation projects using NSGA-II. 2016 IEEE Congress on Evolutionary Computation (CEC), 4208–4213.

Swimming, Basketball, Bicycle Riding, Videography, Traveling, Books, Movies, Languages