

Technion Israel Institute of Technology

**Development of a high-resolution spatial and temporal
electricity demand model in Israel**

Final Report

Primary investigators:

Or Aleksandrowicz, Avigdor Gal

Researchers:

Omri Savir, Gilad Kutiel

**Faculty of Architecture and Town Planning, Faculty of Data and Decision
Sciences**

Heshvan 5787, October 2026

The research was funded by the Ministry of Energy and Infrastructure

Contract number 223-11-056

Abstract

Heating and cooling buildings account for about 30% of Israel's national electricity consumption, and 96.5% of Israeli households now have air conditioning. Yet the data currently used to forecast national demand do not capture how consumption responds to local climate, short-term heat events, building characteristics, or the behaviour of different populations. Urban building energy models have rarely been validated against large real-world datasets, mostly because researchers rarely get access to high-resolution consumption data. This study used large-scale smart-meter data to measure how weather, building characteristics, and user populations shape residential electricity demand. It also aimed to develop models that predict this demand at high spatial and temporal resolution, from the single building up to the national grid.

The study had two parts. *Part I* analysed 15-minute smart-meter readings from the Israel Electric Corporation, aggregated per building. They covered residential buildings in seven cities (Ashdod, Beer Sheva, Hadera, Modiin, Netanya, Rishon LeZion, and Rosh Ha'ayin) in 2023 and 2024, with up to 22,050 buildings and 343,181 meters per year. We linked consumption to hourly Heat Index values from the Israel Meteorological Service, building geometry derived from Survey of Israel data, and the socio-economic cluster of each building's statistical area. From each building's consumption time series, we extracted 782 statistical features, then used them to classify buildings and group them into consumption types. We built cross-building prediction models (linear regression and Explainable Boosting Machine) that predict consumption in buildings not seen during training. We also built per-building Prophet models that forecast summer consumption two months ahead from weather alone, tested with both observed and forecast weather.

Part II dealt with national day-ahead demand forecasting. Assuming under-predicting demand typically costs 5–20 times more than over-predicting it, we trained a compact neural network with asymmetric (pinball) loss functions. For this, we used three years of 5-minute national demand data from the system operator Noga (March 2023 – February 2026) and weather data from three Israel Meteorological Service stations. We compared it with an identical model trained with a symmetric loss and post-hoc cost calibration, as well as Noga's published day-ahead forecasts.

Main findings

Residential consumption closely follows heat stress: the correlation between daily Heat Index and consumption was $r = 0.75$. Summer consumption averaged 1.86 times that of the

transition seasons. Weekend consumption averaged 12–17% higher than on weekdays, and on Fridays it peaked at midday rather than in the evening.

Consumption patterns differed clearly between cities and population groups. A building's consumption "fingerprint" alone identified its city with a weighted F1 score of 0.75–0.79, and identified its neighbourhood's socio-economic cluster less well. Clustering produced seven "construction–consumption" types, three of which cover about 75% of buildings. These types were only partly stable: about half of the buildings that clustered together in 2023 remained together in 2024. Demographics seemed to matter more than construction. Buildings in low socio-economic, mostly ultra-Orthodox neighbourhoods used about as much electricity as buildings in affluent areas.

The cross-building models reached city-level R^2 of 0.98–0.995 and building-level R^2 of 0.46–0.74. Once we included consumption history and the Heat Index, basic building characteristics (floor area, number of floors) added almost nothing. Prophet forecasts reached city-level R^2 of 0.79–0.97 in all but one city-year. Replacing observed weather with weather forecasts reduced accuracy only slightly when the models were trained on forecasts.

In Part II, asymmetric training cut expected procurement cost by 40% at a 5:1 cost ratio and by 73% at 20:1, compared with symmetric training. Post-hoc calibration narrowed this gap but did not close it: the asymmetric models still cost 8% and 18% less. Noga's forecast is accurate ($R^2 \approx 0.98$) but essentially unbiased, so it is not cost-optimal when costs are asymmetric.

Conclusions and recommendations

Accessible data (consumption history and the Heat Index) can support accurate, interpretable demand models at the building, city, and national levels without detailed physical building data. We recommend that system operators and market participants train and evaluate forecasts on cost-sensitive metrics, not symmetric error alone, and that market rules require reporting these metrics. Importantly, the analysed buildings were not a representative sample of the national building stock. Therefore, as smart-meter coverage grows, we recommend creating a representative multi-year dataset linked to building and demographic data. This will enable year-round and heatwave-specific forecasting, prediction for buildings with no consumption history (such as planned neighbourhoods), and national-scale calibration of urban building energy models to support energy planning.

תקציר

חימום וקירור מבנים אחראים לכ־30 אחוזים מצריכת החשמל הלאומית בישראל; במקביל, 96.5 אחוזים ממשקי הבית בישראל מחזיקים כיום במזגן. עם זאת, הנתונים המשמשים כיום לחיזוי הביקוש הלאומי לחשמל אינם מביאים בחשבון את האופן שבו הצריכה מגיבה לאקלים המקומי, לאירועי חום קצרי טווח, למאפייני המבנים ולהתנהגותן של אוכלוסיות שונות. מודלים אנרגטיים עירוניים של בניינים כמעט שלא עברו תיקוף מול מאגרי נתונים גדולים מהעולם האמיתי, בעיקר משום שלחוקרים יש רק לעיתים רחוקות גישה לנתוני צריכה ברזולוציה גבוהה. מחקר זה השתמש בנתוני מונים חכמים בהיקף נרחב כדי לבחון כיצד מזג האוויר, מאפייני הבניינים והרכב אוכלוסיות המשתמשים מעצבים את הביקוש לחשמל בבנייה למגורים. מטרה נוספת של המחקר היתה לפתח מודלים החוזים ביקוש זה ברזולוציה מרחבית וזמנית גבוהה, מרמת המבנה הבודד ועד לרמת רשת החשמל הלאומית.

המחקר כלל שני חלקים. **בחלק הראשון** ניתחנו נתונים רציפים במרווחים של 15 דקות ממונים חכמים של חברת החשמל לישראל המותקנים בבנייני מגורים. מאגר הנתונים שקיבלנו כלל סיכומי צריכה ברמת הבניין הבודד משבע ערים (אשדוד, באר שבע, חדרה, מודיעין, נתניה, ראשון לציון וראש העין) מהשנים 2023 ו־2024, ובסך הכול מ־22,050 בניינים ומ־343,181 מונים. את נתוני הצריכה קישרנו לערכים שעתיים של מדד החום (Heat Index) מהשירות המטאורולוגי, למאפיינים גיאומטריים של הבניינים שהופקו מנתונים שהתקבלו המרכז למיפוי ישראל ולאשכול החברתי-כלכלי של האזור הסטטיסטי שבו נמצא כל בניין. מסדרת הזמן של צריכת החשמל של כל בניין חילצנו 782 מאפיינים סטטיסטיים, ובאמצעותם סיווגנו את הבניינים וקיבצנו אותם לטיפוסי צריכה. פיתחנו מודלים (רגרסיה לינארית ו-Explainable Boosting Machine) לחיזוי הצריכה בבניינים שלא נכללו בשלב האימון. בנוסף, פיתחנו מודלי Prophet לכל בניין בנפרד החוזים את צריכת הקיץ חודשיים מראש על בסיס נתוני מזג אוויר בלבד, ובחנו אותם הן מול נתוני מזג אוויר נמדדים והן מול תחזיות מזג אוויר.

החלק השני של המחקר עסק בחיזוי הביקוש הלאומי ליום המחרת. על בסיס ההנחה שעלויות של חיזוי חסר של הביקוש גבוה בדרך כלל פי 5 עד 20 מעלותו של חיזוי יתר, אימנו רשת נזירות קומפקטית באמצעות פונקציות הפסד אסימטריות (pinball). לשם כך השתמשנו בנתוני ביקוש לאומיים במרווחים של 5 דקות לאורך שלוש שנים שפורסמו על ידי חברת נגה, האחראית על ניהול מערכת החשמל הארצית (מרץ 2023 – פברואר 2026), ובנתוני מזג אוויר משלוש תחנות של השירות המטאורולוגי. את תוצאות המודל השווינו למודל זהה שאומן בפונקציית הפסד סימטרית ולכיוול עלויות בדיעבד של מודל זה, וכן לתחזיות יום המחרת שמפרסמת נגה.

ממצאים עיקריים

צריכת החשמל במגורים עוקבת מקרוב אחר עומס החום: המתאם בין מדד החום היומי לבין הצריכה היה $r = 0.75$. הצריכה בקיץ היתה בממוצע פי 1.86 מהצריכה בעונות המעבר. הצריכה בסופי השבוע היתה בממוצע גבוהה ב־12 עד 17 אחוזים מהצריכה בימי חול, ובימי שישי היא הגיעה לשיאה בצהריים ולא בערב.

דפוסי הצריכה נבדלו באופן ברור בין ערים ובין קבוצות אוכלוסייה. "טביעת האצבע" הצריכתית של בניין איפשרה לזהות את העיר שבה הוא נמצא בציון F1 משוקלל של 0.75–0.79. זיהוי האשכול החברתי-כלכלי של השכונה התאפשר ברמת דיוק נמוכה יותר. ניתוח האשכולות הניב שבעה טיפוסי "בנייה-צריכה", ששלושה מהם כוללים כ־75 אחוז מהבניינים. טיפוסים אלה היו יציבים באופן חלקי בלבד: כמחצית מהבניינים שהשתייכו לאותו אשכול ב־2023 נותרו באותו אשכול גם ב־2024. נראה כי למאפיינים הדמוגרפיים היתה השפעה רבה יותר מאשר למאפייני הבנייה: בניינים בשכונות בעלות דירוג חברתי-כלכלי נמוך, שרובן חרדיות, צרכו חשמל בהיקף דומה לזה של בניינים באזורים אמידים.

מודלי החיזוי הגיעו ל- R^2 של 0.98–0.995 ברמת העיר ול- R^2 של 0.46–0.74 ברמת המבנה. לאחר הכללת היסטוריית הצריכה ומדד החום, מאפייני הבניין הבסיסיים (שטח רצפה, מספר קומות) כמעט שלא תרמו לדיוק החיזוי. תחזיות מודל Prophet הגיעו ל- R^2 של 0.79–0.97 ברמת העיר מלבד בעיר

אחת בשנה אחת מהשתיים. החלפת נתוני מזג האוויר הנמדדים בתחזיות מזג אוויר הפחיתה את הדיוק במידה מועטה בלבד כאשר המודלים אומנו על תחזיות.

בחלק השני של המחקר, האימון האסימטרי הפחית את עלות הרכש הצפויה ב־40 אחוזים ביחס עלויות של 5:1 וב־73 אחוזים ביחס של 20:1, בהשוואה לאימון סימטרי. הכיול בדיעבד צמצם את הפער אך לא סגר אותו: עלות המודלים האסימטריים עדיין הייתה נמוכה ב־8 אחוזים וב־אחוזים בהתאמה. התחזית של נגה מדויקת ($R^2 \approx 0.98$) אך למעשה חסרת הטיה, ולכן אינה אופטימלית מבחינת עלות כאשר העלויות אסימטריות.

מסקנות והמלצות

נתונים זמינים (היסטוריית צריכה ברזולוציה גבוהה ומדד חום) יכולים לשמש בסיס למודלי ביקוש מדויקים וניתנים להסבר ברמת הבניין הבודד, העיר והמדינה, ללא צורך בנתונים פיזיים מפורטים על הבניינים. אנו ממליצים שמנהלי המערכת והשחקנים השונים בשוק יאמנו מודלים ויעריכו תחזיות על סמך מדדים רגישים לעלויות ולא על סמך שגיאה סימטרית בלבד, ושכללי השוק יחייבו דיווח על מדדים אלה. חשוב לציין כי הבניינים שנותחו לא היוו מדגם מייצג של מלאי הבינוי הלאומי. לכן, עם הרחבת הפריסה של מונים חכמים, מומלץ ליצור מאגר נתונים מייצג ורב-שנתי המקושר גם לנתוני בניינים ולנתונים דמוגרפיים. מאגר זה יאפשר חיזוי לאורך כל השנה ובמהלך גלי חום, חיזוי עבור בניינים ללא היסטוריית צריכה (כגון שכונות בתכנון) וכיול בקנה מידה לאומי של מודלים אנרגטיים עירוניים של בניינים לצורך תכנון משק האנרגיה.

Part I: Geography-specific prediction of electricity consumption at the building and city level

1. Introduction

1.1 Scientific and technological background

Although the potential for energy savings in buildings is well recognised, determining their scale is difficult without a detailed analysis of energy consumption data. In European Union countries, where systematic work is being conducted to quantify the energy-saving potential in buildings since the adoption of the Energy Performance of Buildings Directive (EPBD) in 2002, energy consumption forecasts are usually based on the extraction and extrapolation of datasets of sample data that are expected to reliably represent the entire national building stock. Such datasets are used for predicting future trends and for calculating the savings potential of refurbishments, as well as for determining threshold levels for the energy performance of buildings (Ascione et al., 2013; Ballarini et al., 2014; di Staso et al., 2015; Kaden & Kolbe, 2013). Similar work has not yet been done in Israel, even though several official standards include threshold levels for energy performance and the thermal resistance of building envelopes, most notably IS 5280, 5281, and 5282.

Relying on models, standards, and simulations instead of real consumption data at the individual building level may be a source of significant errors, partly due to the difficulty of realistically quantifying the energy demands of different types of users (Galvin, 2014; Malik et al., 2022; Mathew et al., 2015; G.-S. Olivia & T. A. Christopher, 2015; Prativiera et al., 2022; Sunikka-Blank & Galvin, 2012). While energy providers maintain detailed datasets on energy consumption, they almost always do so without linking consumption data to detailed data on the buildings where the energy is consumed and on the demographic and socio-economic characteristics of their occupants. Therefore, while the potential of using "big data" for large-scale and detailed analysis (Few et al., 2023; McKenna et al., 2022) or at least mapping (Howland, 2013; Vecchi et al., 2023) of actual energy consumption in buildings does exist, but how it is affected by different variables at different scales (a city district, a neighbourhood, a city block, or a single building) is rarely explored.

1.2 Current state of research

In the last decade, most of the studies that have attempted to analyse residential energy consumption focused on households as their reference point, not on buildings (Andersen et al., 2021). This approach to energy consumption analysis may have resulted from a focus on defining typical user profiles as the basis for developing a national building energy prediction, but may also reflect researchers' indifference to the effects of construction technology and building location on energy consumption, as well as to energy management and economics. As a result, these studies usually relied on data from representative samples of hundreds or thousands of households rather than comprehensive datasets covering entire city blocks, neighbourhoods, or cities, which prevented systematic consideration of buildings' physical properties and locations as independent variables.

In contemporary research literature, few documented attempts have mapped real consumption data for buildings, city blocks, or neighbourhoods at an urban scale, and none used the data to model predicted energy consumption. One of the first studies to employ such a methodology was conducted in 2011 by researchers at MIT (Kolter & Ferreira, 2011). The study produced an energy map that presented the monthly energy consumption of 6,500 buildings in Cambridge, Massachusetts. The data were compared with energy consumption forecasts produced by a statistical probability procedure, without considering the buildings' physical and thermal characteristics or calculating expected energy consumption using simulation tools. A somewhat similar approach was adopted in a 2012 study by researchers at Columbia University, who created a map showing annual energy consumption per block in New York City, based on real-life consumption figures (Howard et al., 2012). They extracted heating and cooling data from general consumption data based on typical consumption surveys and the building's use classification. The results were used to estimate energy demand based on building use and block area.

In 2013, researchers at UCLA mapped the monthly energy consumption of urban blocks in Los Angeles, cross-referencing real consumption data with information on land use, income level, building age, and block area (Pincetl et al., 2013). Researchers also examined a possible correlation between building age and energy consumption by analysing monthly

gas consumption for about 1,350 buildings over two quarters in Basel (Aksoezen et al., 2015) and by examining district heating meters in 15,000 buildings in Stockholm using hourly consumption data (Shahrokni et al., 2014). Other studies from the last five years used relatively small datasets of several hundred smart meters mainly to relate energy consumption to household characteristics (Chévez & Barbero, 2023; Lazzari et al., 2022; McKenna et al., 2022), but also to building types (Westermann et al., 2020). Recently, a study in Israel attempted to correlate electricity consumption with apartment size using monthly consumption data from about 800 households (Bugin et al., 2022).

These studies, while demonstrating the potential of analysing real-life, high-resolution energy consumption data to examine its relationship with building and user characteristics, implicitly highlighted the challenge of data availability. In most countries, researchers still rarely have access to large datasets of user consumption figures at high temporal resolution. One notable exception is the Danish smart-meter dataset, which contains data on about 700,000 anonymised households and is managed by Statistics Denmark. This open and unique dataset combines high-resolution household consumption data collected by EnergiNet, the Danish national transmission system operator for electricity and natural gas, and household statistics collected by Statistics Denmark. In recent years, this constantly updating dataset has been used by various researchers primarily to explore the effects of socio-economic characteristics on peak energy demand and typical usage profiles (Andersen et al., 2021; Gunkel et al., 2023), yet these studies did not attempt to relate consumption to building characteristics or location.

A smaller, yet relatively similar, dataset consisting of two years of smart-meter data from about 110,000 households was recently created in Uruguay by researchers from the Universidad de la República (Chavat et al., 2022) and was used to define "socio-technical" urban clusters demonstrating similar socio-economic characteristics, apartment sizes, building typologies, and consumption trends (Chévez & Barbero, 2023). Nevertheless, the study's analytical resolution was no higher than the census tract level and did not include an in-depth analysis of building characteristics; instead, it focused on the tract's average socio-economic status.

1.3 Existing knowledge gaps

Although the potential for energy savings in buildings is widely acknowledged, determining their magnitude is difficult without a detailed analysis of energy consumption data and a better understanding of what affects it. In recent years, considerable efforts have been made to develop tools and models, commonly referred to as Urban Building Energy Models (UBEM), for predicting energy consumption in buildings in large urban clusters based on construction inventory characteristics (Ferrando et al., 2020; Gassar & Cha, 2020; Johari et al., 2020). Yet none of these models has gone through an extensive process of validation vis-à-vis large real-life datasets of energy consumption in buildings (Abbasabadi & Ashayeri, 2019; Hong et al., 2020).

Relying on unvalidated "synthetic" building energy models can be a source of significant errors, in part due to the difficulty of realistically defining user behaviour patterns or the type of heating and cooling systems used (de Wilde, 2021; Galvin, 2014; Mathew et al., 2015; G. S. Olivia & T. A. Christopher, 2015; Sunikka-Blank & Galvin, 2012). In Israel, where the percentage of households with air conditioning systems rocketed between 2004 and 2023 from 65.8% to 96.5% (ranging from 89% in the lower decile to 98.3% in the upper decile) according to data from the Central Bureau of Statistics, an error originating from a lack of data on a household's possession of an air conditioning system is relatively negligible. At the same time, the gap between estimated and actual consumption could be particularly large because air conditioning is almost always at hand, yet how different population groups use it remains unknown. Given the significant differences in living standards, cultural habits, and air-conditioning usage preferences across populations in different urban areas, there is a clear need to identify and quantify these differences to develop realistic tools for generating electricity consumption predictions at high spatial and temporal resolutions.

1.4 Research objective

Reducing energy consumption has been a declared national goal in Israel since 2008. Electricity use in buildings is an essential component of overall national energy consumption and, therefore, a key factor in meeting energy-savings goals, especially as

electric indoor heating and cooling steadily rise. According to various estimates, heating and cooling buildings in Israel account for about 40% of residential and about 60% of commercial annual electricity consumption, totalling about 30% of total national electricity consumption (Bet-Hazavdi & Drori, 2010). Increased electricity consumption in buildings can reflect rising standards of living, reduced tolerance of indoor heat among building occupants, poor thermal performance of building envelopes, and increased building cooling loads driven by global warming. However, we still have little empirical evidence to quantify the share of each variable in overall electricity consumption. Given the significant differences in living standards, cultural habits, and air-conditioning usage preferences across urban populations, we need to identify and quantify these differences to develop realistic tools for predicting electricity consumption at high spatial and temporal resolutions.

Recently, however, due to developments in the methods of measuring and storing data on energy consumption and the increasing distribution of geographic information systems, it is easier to use "big data" for large-scale mapping and analysis of actual energy consumption in buildings and to exploit it for defining recurrent or abnormal patterns of electricity usage in buildings. With these developments, it is also possible to produce much more accurate models than those currently available for predicting building energy consumption across scales of analysis ranging from a single building to an entire urban area. The main stumbling block in this context was, and remains, access to high-resolution raw data on actual building energy consumption.

2. Methods

This part of the study is based on large-scale smart meter data gathered by the Israel Electric Corporation (IEC). By the end of 2023, the IEC had installed 660,000 smart meters (Israel Electric Corporation, 2023), and this number is expected to rise in the next few years, as the IEC intends to replace the entire old metering stock for Israeli electricity end-users. These meters collect consumption data at 15-minute intervals and can therefore be used to accurately evaluate consumption levels against short-term fluctuations in outdoor weather

conditions. For this study, we analysed smart-meter data covering at least 12 consecutive months from ~225,000 meters across 7 Israeli cities to build a comprehensive prediction model. Access to such a large smart-meter dataset is rare for researchers, not only in Israel but also globally (Wang et al., 2026).

By analysing long-term and short-term consumption fluctuations against meteorological data from the Israel Meteorological Service (IMS), and building characteristics data from the Survey of Israel's GIS building footprint data layer, it was possible to produce a solid quantitative base for developing a model for predicting energy consumption in buildings, from the individual building level up to the urban level, while considering certain aspects of the building stock, climatic and meteorological data, and electricity usage profiles.

2.1 Data

2.1.1 Electricity consumption data

Residential electricity consumption data were obtained from the IEC and consisted of readings taken at 15-minute intervals, aggregated per building to secure user anonymity, from 7 Israeli cities: Ashdod, Beer Sheva, Hadera, Modiin, Netanya, Rishon LeZion, and Rosh Ha'ayin. The data represented all buildings with more than 3 residential units and smart meters from each city with a significant rate of smart meter installations at the beginning of 2024, which led to uneven geographic distribution of the selected buildings, either within the cities or across Israel. In addition to total consumption, the dataset included a postal address and the number of installed smart meters for each building, which enabled us to verify that the data came only from residential units. We received 2 full annual datasets: the complete year 2023 (14,863 buildings, 264,769 meters) and the complete year 2024 (22,050 buildings, 343,181 meters). We also used summer 2022 data from Modiin (June–September 2022), covering 1,904 buildings, for initial model development. Table 1 and Table 2 summarise the number of buildings and meters available, respectively, per city and per year.

Table 1: Number of buildings per city and year

City	2023	2024
Ashdod	3,997	4,413
Beer Sheva	1,264	2,243
Hadera	1,677	2,196
Modiin	2,610	2,958
Netanya	1,412	3,300
Rishon LeZion	3,431	5,755
Rosh Ha'ayin	472	1,185
Total	14,863	22,050

Table 2: Number of meters per city and year

City	2023	2024
Ashdod	72,046	75,378
Beer Sheva	25,351	36,106
Hadera	28,781	31,097
Modiin	25,865	29,370
Netanya	37,473	62,171
Rishon LeZion	62,801	91,467
Rosh Ha'ayin	12,452	17,592
Total	264,769	343,181

2.1.2 Meteorological data

Meteorological monitoring data were collected from the IMS for 2023 and 2024 at a raw temporal resolution of 10 minutes, then aggregated to hourly values. The data originate from 5 meteorological stations: Ashdod (used for Ashdod), Beer Sheva (used for Beer Sheva), Nachshon (used for Modiin), Hadera (used for Netanya and Hadera), and Beit Dagan (used for Rosh Ha'ayin and Rishon LeZion). Based on temperature and relative humidity readings, hourly Heat Index values were computed for each city following a commonly used calculation method described in Rothfusz (1990).

The Heat Index serves as the primary environmental variable in all prediction models developed in this study because it captures the combined effect of air temperature and relative humidity on perceived heat stress and can therefore better predict heat stress intensity and subsequent cooling loads and air conditioning use than air temperature alone. We also computed lagged averages of the Heat Index over windows of up to 7 prior days to capture delayed thermal effects. In addition to monitored air temperature and relative humidity readings, we received short-term weather forecast data from the IMS for 2023 from the same 5 meteorological stations to evaluate prediction accuracy based on predicted weather conditions. This dataset consisted of hourly forecasts of air temperature and relative humidity up to 90 hours ahead, issued once daily at 00:00.

2.1.3 Building data

We located each building for which we received electricity consumption data based on the address in the IEC dataset. This enabled us to link each building to socio-economic statistics from Bureau of Statistics surveys and statistical zone definitions, supporting spatial understanding of building distribution within each city.

In addition, we geolocated addresses to extract several building characteristics: building height, building volume, total gross building floor area, and number of floors. We did this by processing 2 geographic datasets for the 7 cities received from the Survey of Israel: building footprints and a Digital Surface Model (DSM). Because total floor area is not a commonly available building attribute, we calculated it by combining building height data

derived from the 50 cm/pixel DSM with the building footprint polygons. Assuming a typical floor height of 3 m, we estimated the number of floors per square metre of footprint, yielding each building's total gross floor area. However, in retrospect, building characteristics contributed negligibly to model accuracy once we included consumption data and the Heat Index.

2.2 Data processing and manipulation

2.2.1 Aggregation to hourly resolution

The raw IEC electricity consumption readings are recorded at 15-minute intervals. For each building-day, we summed the four quarter-hour readings within each clock hour (HH:00, HH:15, HH:30, HH:45) to produce a single hourly consumption value, yielding a long-format table with one row per building-hour timestamp. We aligned the resulting hourly consumption series with the IMS meteorological data by matching city and timestamp fields. We then divided a building's absolute hourly consumption [kWh] by the number of meters affiliated with that building in the original IEC dataset to normalise consumption levels. To efficiently remove outliers, we also normalised consumption by building gross floor area (Section 2.1.3), producing consumption levels by area [kWh/m²]; since this latter normalisation type involved an additional level of uncertainty introduced by the method used to extract building gross floor area, we did not use it for creating the prediction models.

2.2.2 Missing data removal

We identified and removed missing values in both the meteorological and consumption datasets before analysis and enforced series completeness per building and per season. The year was divided into 5 seasons: winter 1 (January–March), spring (April–May), summer (June–September), autumn (October–November), and winter 2 (December). Winter was split into 2 separate periods (winter 1 and winter 2) because the data cover a calendar year, and a single continuous winter season would span across the year boundary. Each season has its own expected record count (e.g., summer = June–September = 122 days × 24 hours = 2,928 hourly records per building). Buildings whose actual record count did not exactly match the expected count for a given season were excluded from that season's analysis.

For 2023, January and February showed a disproportionate number of missing and anomalous readings and were therefore excluded from the classification analyses (2.3.8); in these cases, analyses used data from March through December only.

2.2.3 Outlier removal

We applied outlier removal in two sequential steps. First, we removed buildings whose mean hourly consumption exceeded an absolute threshold of 90 kWh/m², as these correspond to non-residential meters or metering anomalies. Second, among the remaining buildings, we also excluded those whose per-building mean consumption exceeded the city-wide mean by more than 3 standard deviations. A subsequent address-level inspection confirmed that all removed buildings were commercial or industrial, not residential. Table 3 and Table 4 summarise the number of buildings and meters, respectively, remaining after preprocessing per city and per year (summer only).

Table 3: Number of buildings remaining after preprocessing with available summer data, per city and year

City	2023	2024
Ashdod	2,491	4,128
Beer Sheva	765	1,685
Hadera	1,559	2,054
Modiin	2,321	2,669
Netanya	1,012	1,706
Rishon LeZion	1,876	4,871
Rosh Ha'ayin	346	477
Total	10,370	17,590

Table 4: Number of meters remaining after preprocessing with available summer data, per city and year

City	2023	2024
Ashdod	64,034	66,203
Beer Sheva	21,045	31,721
Hadera	24,836	25,510
Modiin	22,479	24,531
Netanya	30,773	48,701
Rishon LeZion	49,799	76,938
Rosh Ha'ayin	11,567	11,847
Total	224,533	285,451

2.3 Exploratory data analysis

2.3.1 Relationship between weather and electricity consumption

Ambient thermal conditions may strongly drive electricity consumption in residential buildings, primarily through air conditioning in summer and heating in winter. This relationship is further amplified by the very high penetration of air conditioning in Israeli households. Figure 1 shows the daily average normalised consumption across all buildings in the dataset throughout 2024. A sharp, pronounced elevation appears during the summer months (June–September), driven by cooling demand. A more moderate secondary elevation appears in winter (December–March), consistent with heating demand. Day-to-day variability within the summer peak is substantial, reflecting thermal-load fluctuations driven by heat events. However, the sharp increase in summer electricity consumption may also relate to school vacations, which may increase residential occupancy hours.

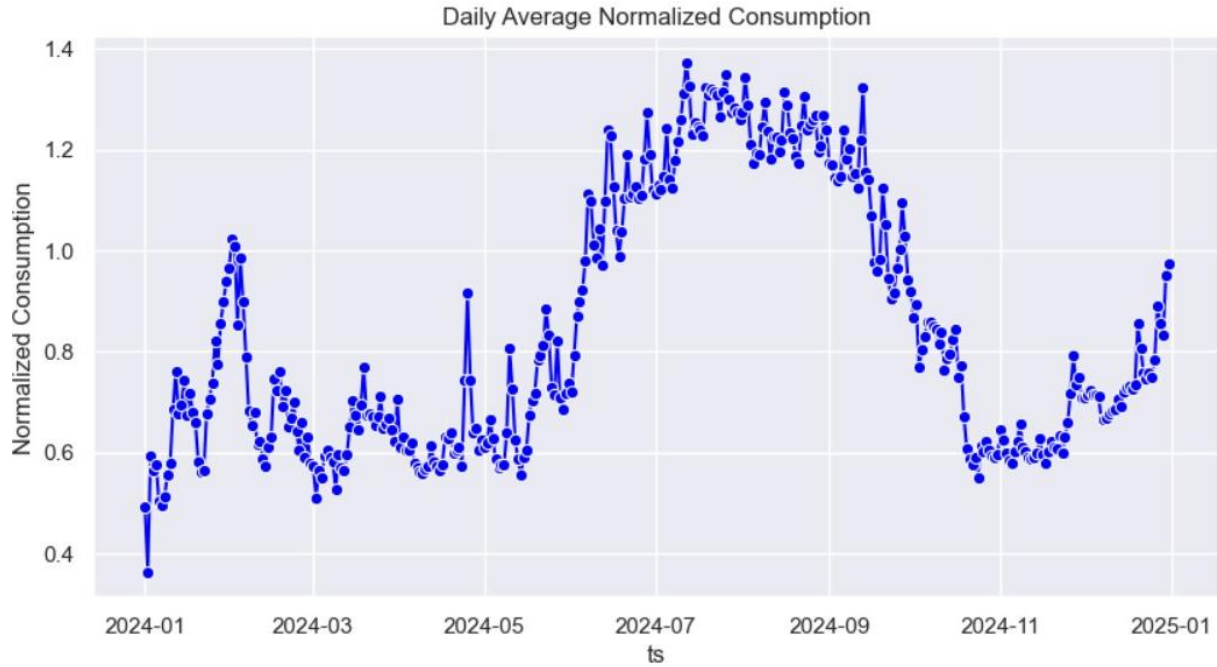


Figure 1: Daily change in average electricity consumption across all dataset buildings, 2024. Consumption values are average hourly kWh per meter (apartment).

Figure 2 quantifies this relationship directly, presenting a scatter plot of daily mean Heat Index against daily mean electricity consumption, pooled across all dataset buildings and days in 2024. The overall Pearson correlation is $r = 0.75$, confirming a strong positive association between thermal load and electricity demand. The relationship holds across the full range of Heat Index values, with consumption variance increasing at higher values, reflecting heterogeneity in building-level cooling capacity and occupancy patterns. At the lower end of the Heat Index range ($HI < 15$), consumption also shows a moderate upward trend, consistent with heating demand during cold periods.

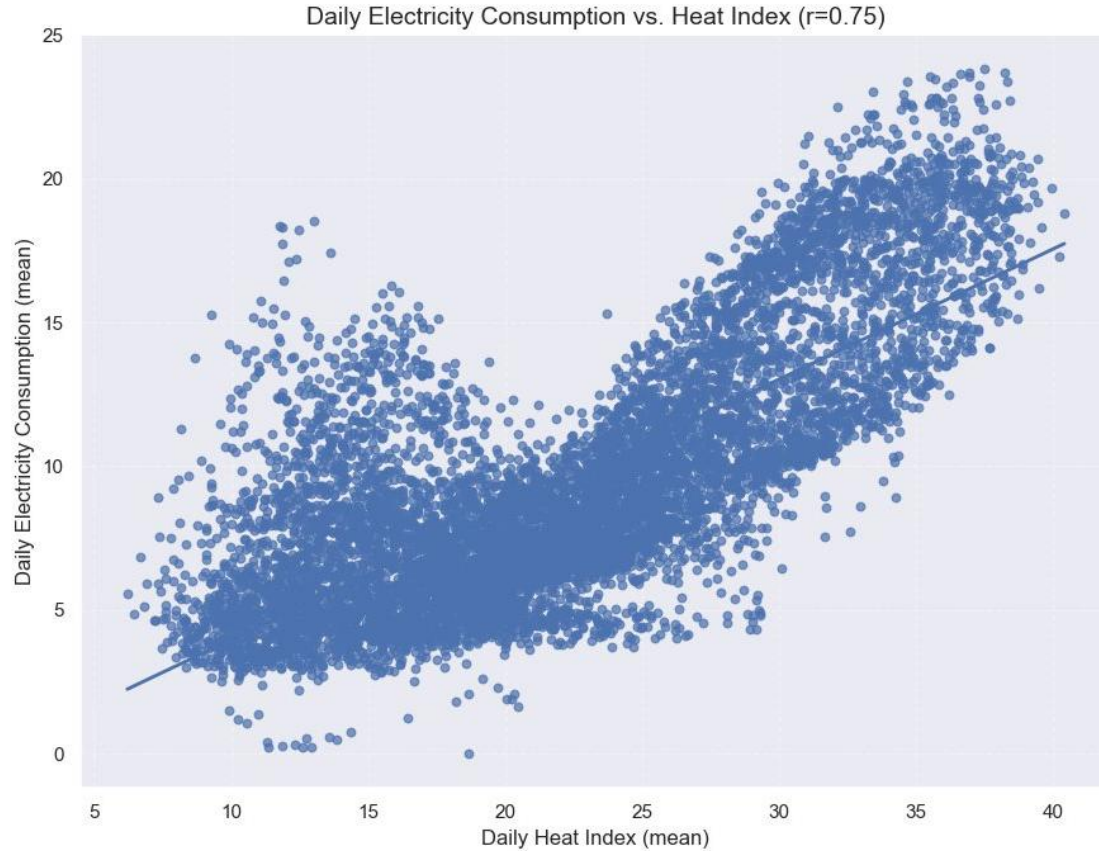


Figure 2: Scatter plot of daily mean Heat Index vs. daily mean electricity consumption (2024), with overall regression trend line ($r = 0.75$). Consumption values are in kWh per meter (apartment).

Further evidence of the strong relationship between thermal load and electricity consumption comes from the summer ratio, defined as the ratio of a building's average hourly consumption during the summer months (June–September) to its average during the transition seasons (April–May and October–November). Figure 3 shows the distribution of this ratio across all buildings. The mean ratio is 1.861. The distribution is approximately bell-shaped and centred around 1.8–1.9, with very few buildings below 1.0, confirming that the summer elevation is a near-universal feature of residential consumption.

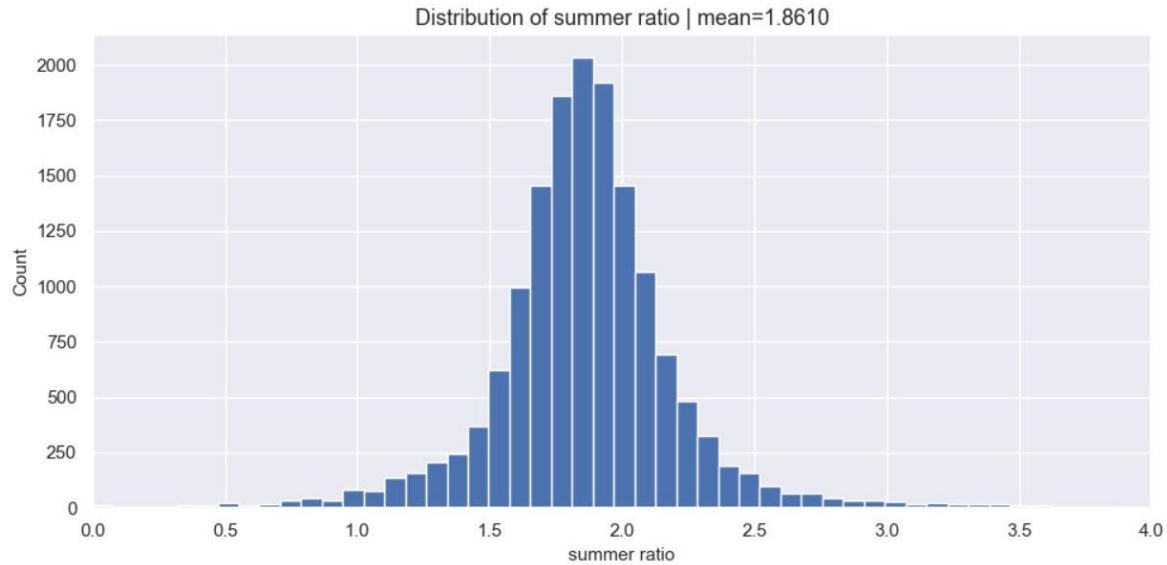


Figure 3: Distribution of the summer ratio across all buildings (ratio of summer to transition-season average hourly consumption; mean = 1.861).

2.3.2 Weekly and intraday consumption patterns

Electricity consumption varies systematically by day of the week, driven primarily by occupancy patterns. Weekend days (Friday and Saturday in Israel) show markedly higher consumption than weekdays, as residents spend more time at home and therefore use appliances, lighting, and climate control more. This pattern is consistent across both seasons and hours of the day.

Figure 4 shows the average normalised consumption by hour of day and day of the week during the warm season. Consumption is elevated on days 6 and 7 (Friday and Saturday) relative to days 1–5 (weekdays) across all hours, with the Friday afternoon block reaching the highest values in the dataset. Figure 5 presents the average consumption by day of the week and month across the full year, confirming that the weekend elevation is a stable feature throughout the year and is particularly pronounced during the summer peak (months 6–9).

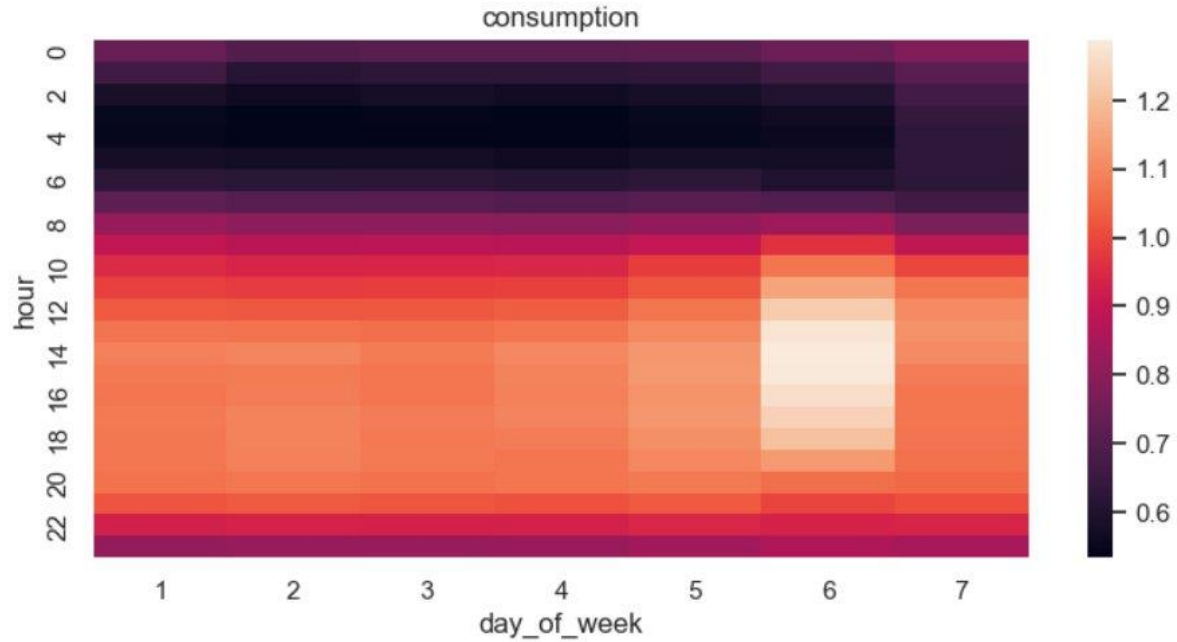


Figure 4: Heatmap of average normalised consumption by hour of day and day of week during the warm season (1 = Sunday, 7 = Saturday) for the entire 2023-2024 dataset. Consumption values (right-hand Y-axis) are average hourly kWh per meter (apartment).

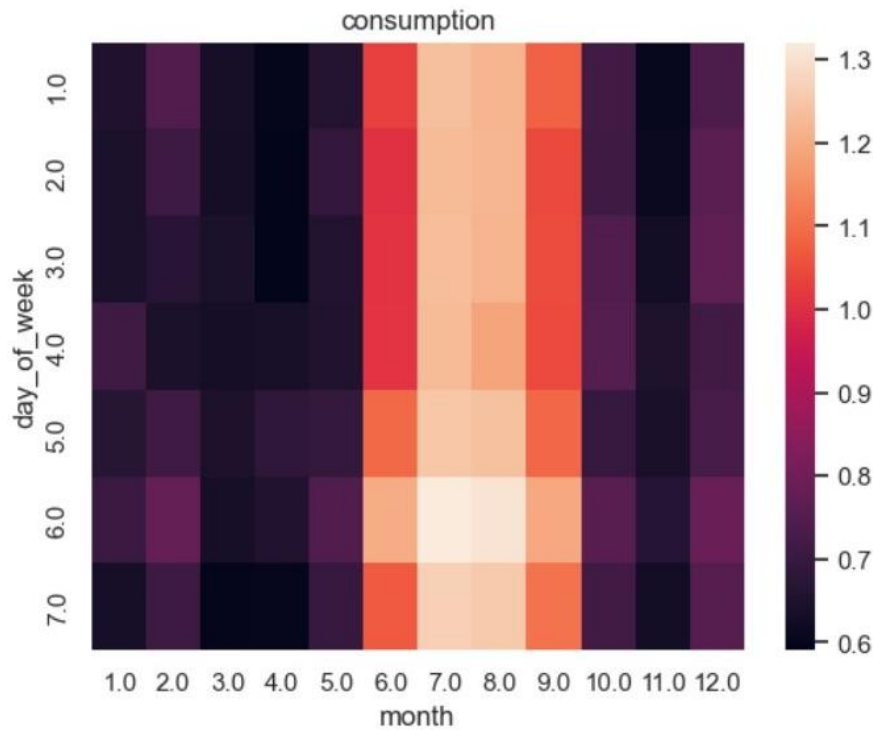


Figure 5: Heatmap of average consumption by day of week and month across the full year (1 = Sunday, 7 = Saturday) for the entire 2023-2024 dataset. Consumption values (right-hand Y-axis) are average hourly kWh per meter (apartment).

Weekend consumption elevation is also quantifiable at the building level. The Friday ratio and Saturday ratio are defined as the ratio of a building's average hourly consumption on Fridays and Saturdays to its average weekday consumption. Figure 6 and Figure 7 show the distributions of these ratios. The mean Friday ratio is 1.165, and the mean Saturday ratio is 1.123. Both distributions are approximately unimodal and right-skewed, confirming that most buildings show higher weekend consumption, although a notable minority has ratios below 1.0, consistent with the city-level reversal observed for Beer Sheva, Netanya, and Rosh Ha'ayin.

Beyond overall consumption levels, the intraday shape of the consumption profile also differs between day types. Figure 8 presents polar coordinate plots of average hourly consumption for Friday, Saturday, and weekdays. On weekdays, the dominant peak occurs in the late afternoon and evening (approximately 16:00–20:00), reflecting residents returning from daily activities. In contrast, on weekends (Friday and Saturday), the consumption profile follows the thermal load more closely: as ambient temperatures rise through the morning, consumption climbs steadily and peaks around midday (approximately 12:00–14:00) on Fridays. This may reflect behaviour in which residents stay home all day and use air conditioning continuously, which responds to the rising thermal load with a lag of about 2 hours, as indoor temperatures take time to rise in response to increased outdoor heat. In all day types, nighttime consumption (approximately 00:00–06:00) is substantially lower than daytime consumption.

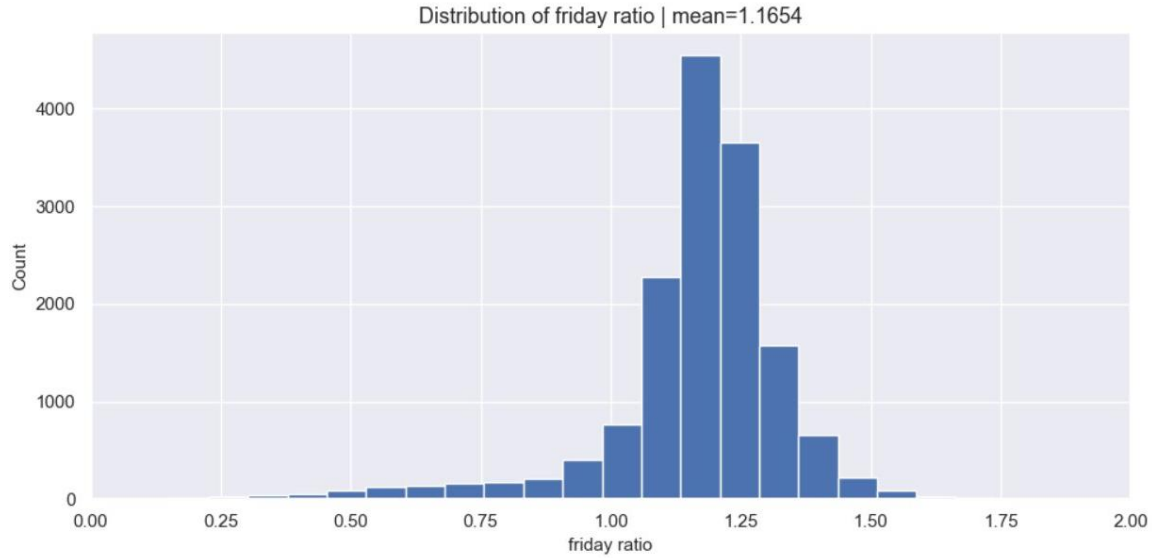


Figure 6: Distribution of the Friday ratio across all buildings for the entire 2023-2024 dataset (ratio of Friday to regular-weekday average hourly consumption; mean = 1.165).

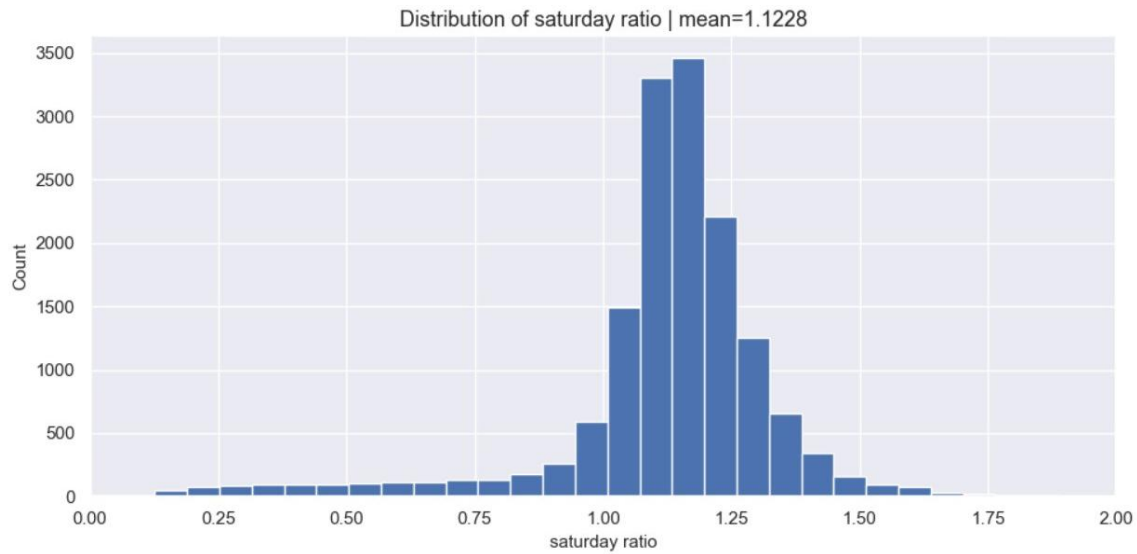


Figure 7: Distribution of the Saturday ratio across all buildings for the entire 2023-2024 dataset (ratio of Saturday to regular-weekday average hourly consumption; mean = 1.123).

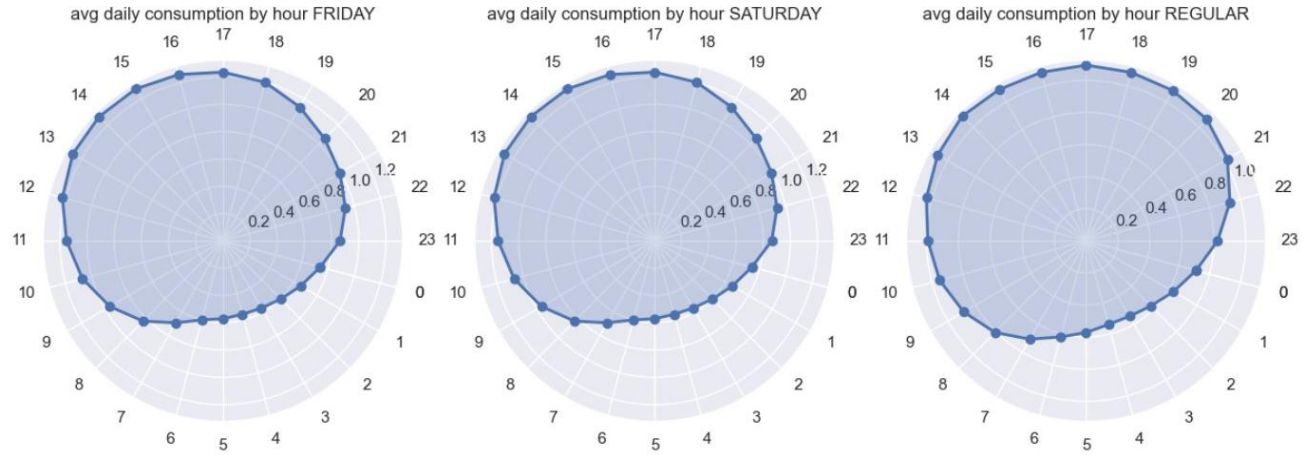


Figure 8: Polar coordinate plots of average hourly electricity consumption by day type for the entire 2023-2024 dataset: Friday (left), Saturday (centre), and regular weekday (right). Consumption values are average hourly kWh per meter (apartment).

Figure 9 and Figure 10 illustrate the intraday relationship between the Heat Index and electricity consumption using hourly 2022 data from Modiin. Figure 9 covers a 2-week period (July 17–30, 2022) and shows the strongly cyclic structure of both series, with each daily cycle of the Heat Index accompanied by a corresponding cycle in consumption. Figure 10 zooms into 3 consecutive days (Thursday, July 28 to Saturday, July 30) and makes the lag relationship particularly clear: on Friday and Saturday (weekend), the consumption peak follows the Heat Index peak with a delay of approximately 2 hours, reflecting the time it takes for indoor temperatures to respond to outdoor thermal load. On Thursday (a weekday), the consumption peak appears in the late afternoon, possibly consistent with residents returning home rather than tracking the Heat Index directly. This provides additional individual-building-level evidence for the weekday/weekend behavioural difference described above.

The intraday consumption profile is further characterised by the hour of peak consumption and the hour of minimum consumption, computed per building. Figure 11 shows that the distribution of peak consumption hour is strongly concentrated in the afternoon, with a mean of 16:00–17:00 and the large majority of buildings peaking between 15:00 and 20:00. Figure 12 shows the distribution of minimum consumption hour: the mean is around 05:00–06:00, with most buildings reaching their daily minimum in the early pre-

dawn hours (02:00–06:00). The distribution has a long right tail extending into the morning and early midday hours, reflecting buildings with uncommon occupancy or operating patterns.

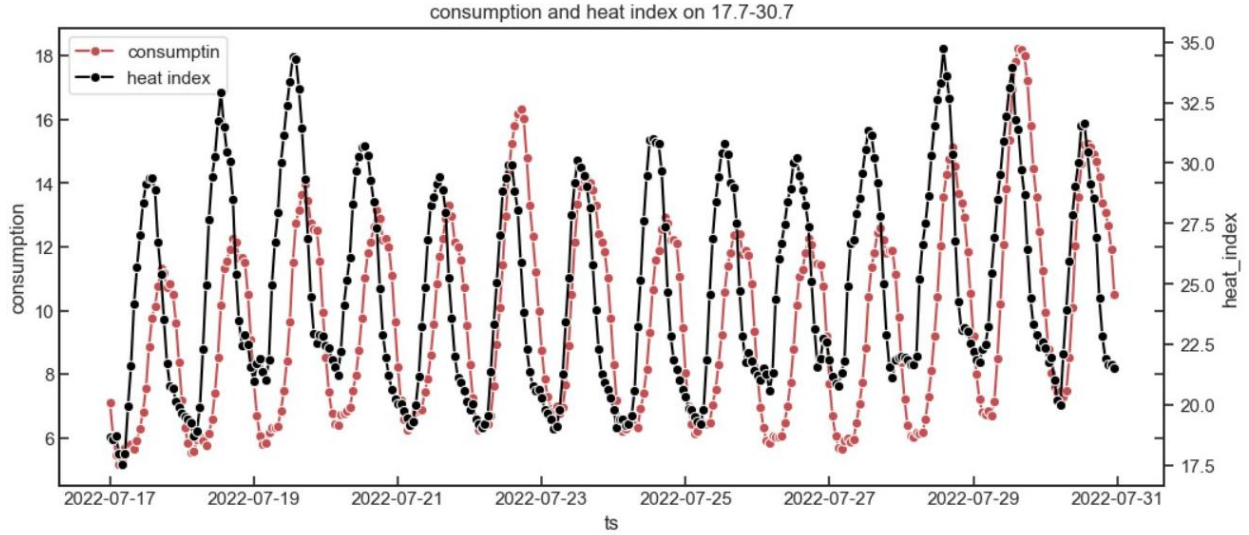


Figure 9: Hourly citywide mean consumption and Heat Index, Modiin, 17–30 July 2022. The daily thermal and consumption cycles are clearly co-varying. Consumption values are average hourly kWh per building.

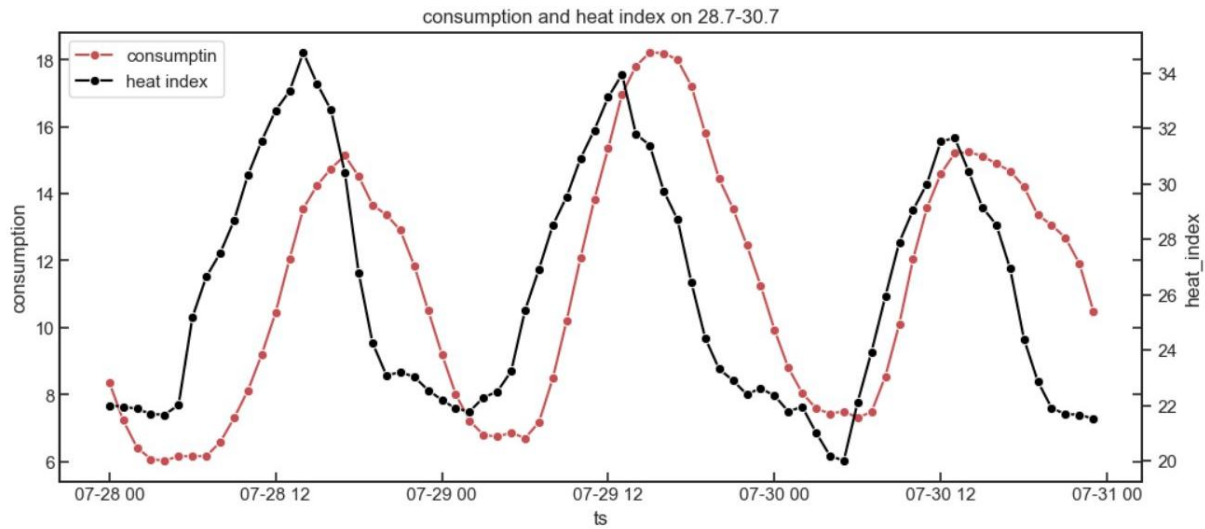


Figure 10: Hourly citywide mean consumption and Heat Index, Modiin, 28–30 July 2022 (Thursday–Saturday). On Friday and Saturday, the consumption peak lags the Heat Index peak by approximately 2 hours; on Thursday, the peak occurs later in the afternoon, consistent with residents returning from work. Consumption values are average hourly kWh per building.

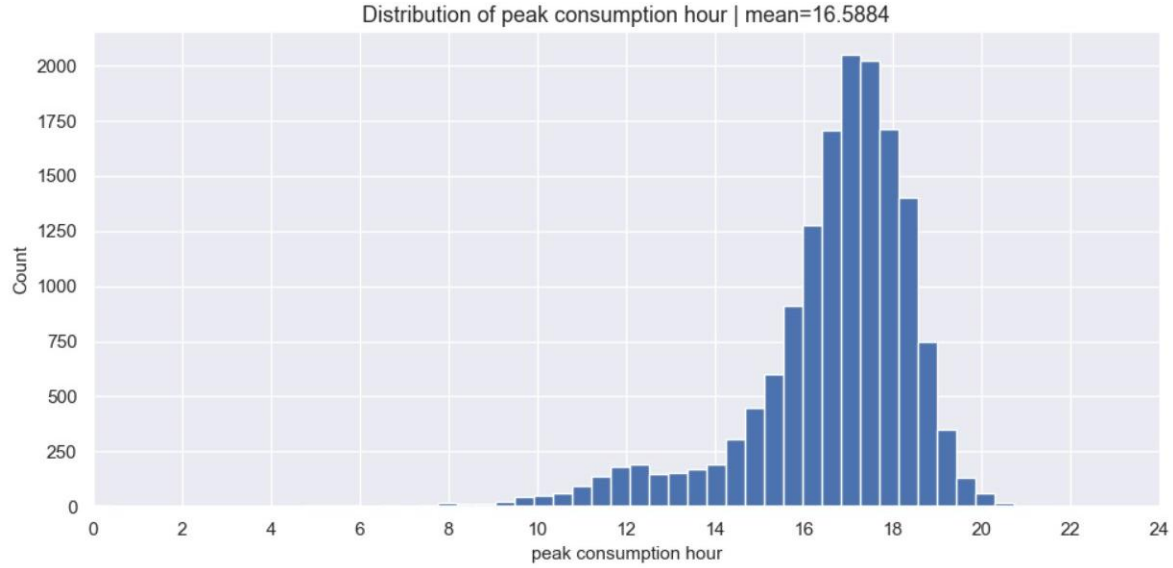


Figure 11: Distribution of the peak consumption hour across all buildings for the entire 2023-2024 dataset (mean = 16.59).

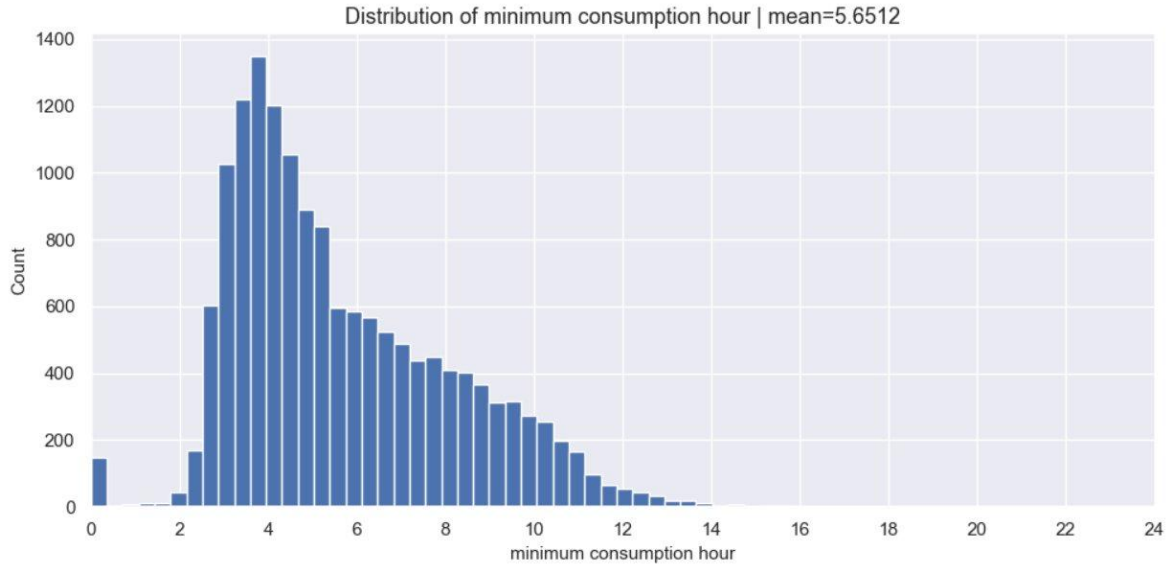


Figure 12: Distribution of the minimum consumption hour across all buildings for the entire 2023-2024 dataset (mean = 5.65).

2.3.3 Inter-city consumption patterns

The seven cities in the dataset show distinct electricity consumption profiles, reflecting interrelations among differences in building characteristics, socioeconomic characteristics, climatic characteristics, and consumer behaviour. Figure 13 presents the distribution of average hourly normalised consumption per meter (apartment) for each city.

Rosh Ha'ayin has the highest median consumption and a relatively narrow interquartile range. At the other end of the spectrum, Beer Sheva and Netanya display low medians but wide interquartile ranges, indicating high within-city variability. Modiin occupies an intermediate position, with a relatively high median and a narrow spread; its low within-city variability likely reflects a relatively homogeneous population compared with the other cities in the dataset. Ashdod, Hadera, and Rishon LeZion cluster closely together at moderate consumption levels.

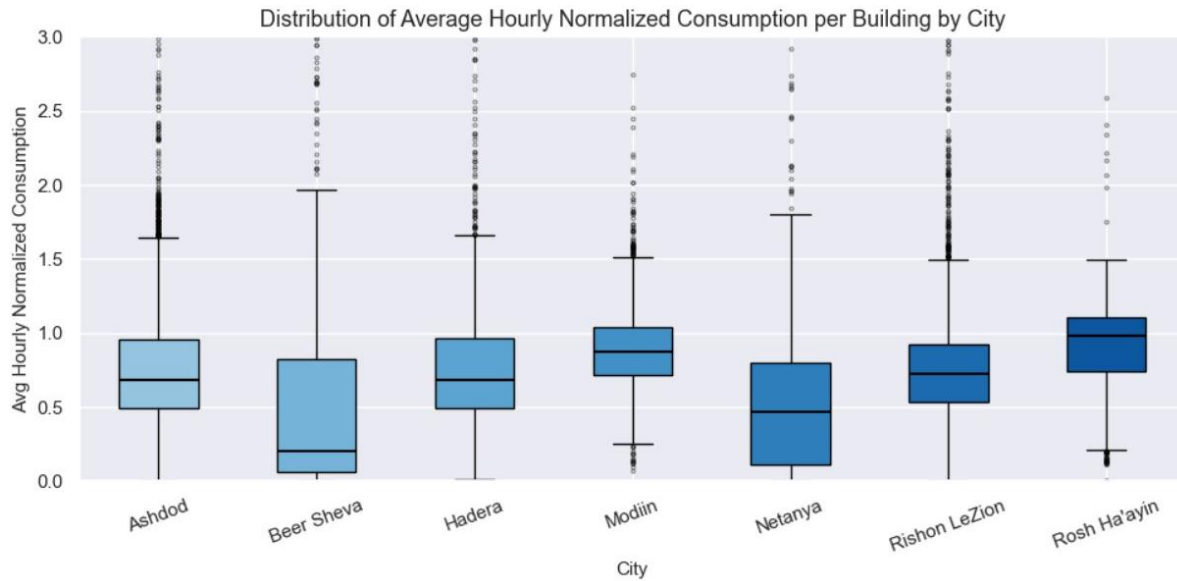


Figure 13: Distribution of average hourly normalised consumption per meter by city. Boxes show the interquartile range; whiskers extend to the 5th and 95th percentile. Consumption figures represent average hourly kWh per meter (apartment).

The heatmap in Figure 14 shows average hourly meter consumption by city and month, revealing that all cities follow the same broad seasonal structure described in Section 2.3.1: a pronounced summer peak from June through September and lower consumption in winter and the transition seasons. The summer peak is most intense for Rosh Ha'ayin, consistent with its elevated building-level consumption observed in Figure 13. Beer Sheva and Netanya remain comparatively low across all months.

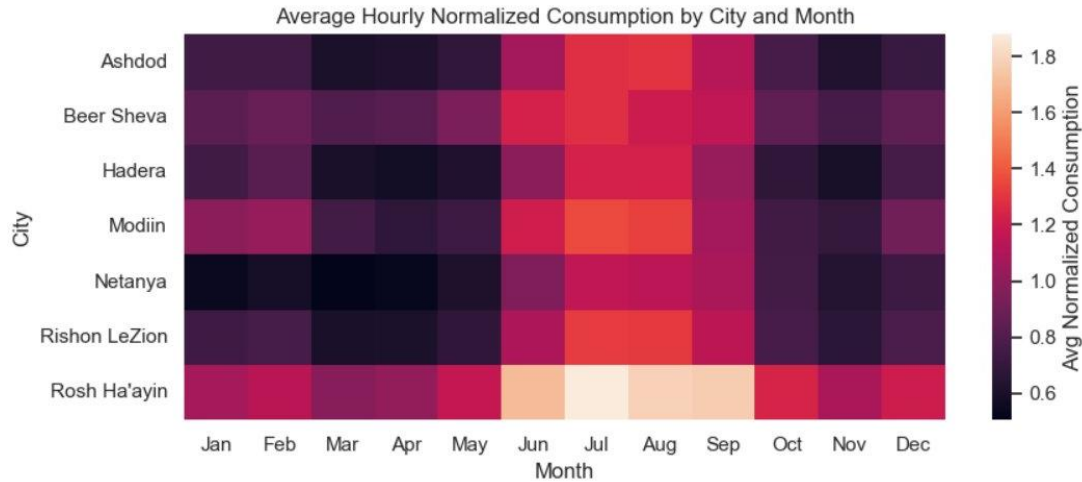


Figure 14: Heatmap of average hourly normalised consumption by city and month (2024). Consumption figures represent average hourly kWh per meter (apartment).

Figure 15 disaggregates consumption by city and day of week. For most cities, Friday and Saturday (days 6 and 7) show higher consumption than weekdays, consistent with the occupancy-driven pattern described in Section 2.3.2. However, Netanya, Beer Sheva, and Rosh Ha'ayin show the opposite pattern: Friday and Saturday consumption is lower than on weekdays. While this reversal may reflect socioeconomic, demographic, and religious differences between these cities and the rest, we cannot definitively conclude what drives these differences because we do not know whether the dataset buildings constitute a representative sample of each city's entire building stock.

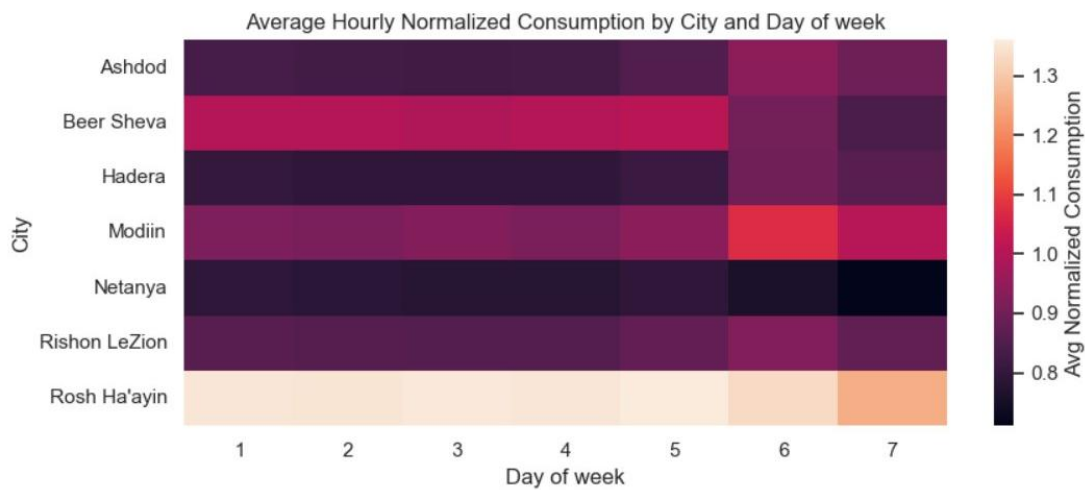


Figure 15: Heatmap of average hourly normalised consumption by city and day of week (1 = Sunday, 7 = Saturday). Consumption figures represent average hourly kWh per meter (apartment).

The intraday consumption profile also varies across cities, as shown in Figure 16. Beer Sheva stands out most clearly, with elevated consumption in the morning and midday hours (approximately 08:00–15:00) relative to other cities, and a sharper decline in the evening. Netanya and Rosh Ha’ayin show a similar, though less pronounced, morning-heavy consumption pattern. This pattern may reflect differences in occupancy schedules. In contrast, Modiin and Ashdod display a flatter intraday profile with consumption remaining elevated into the late afternoon and evening. All cities share the same pronounced nighttime minimum between 00:00 and 06:00.

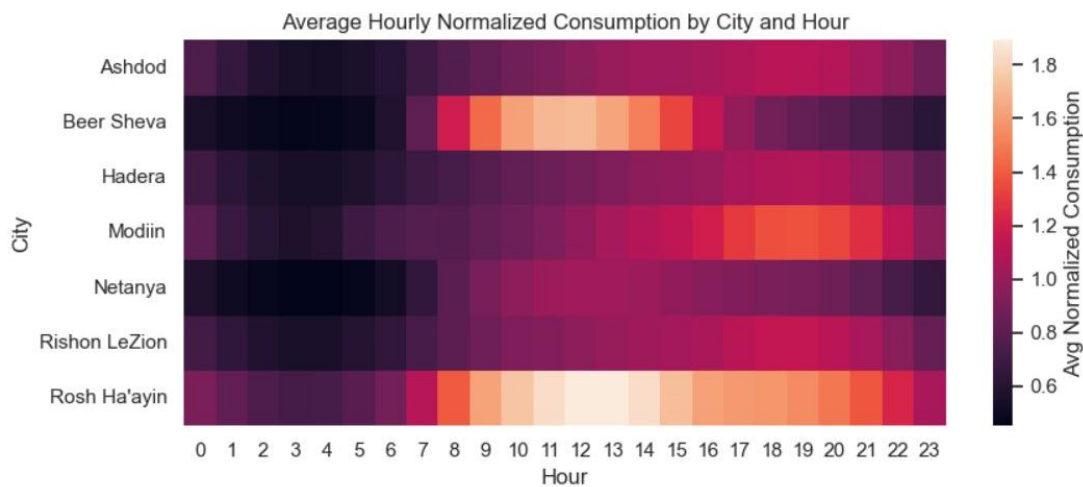


Figure 16: Heatmap of average hourly normalised consumption by city and hour of day (2024). Consumption figures represent average hourly kWh per meter (apartment).

Figure 17 and Figure 18 show the distribution of peak and minimum consumption hours by city. Most cities are broadly similar: peak consumption concentrates in the afternoon hours (16:00–18:00) and minimum consumption in the early morning (04:00–06:00), consistent with the population-wide patterns described in Section 2.3.2. Beer Sheva is a notable exception: it has a substantially larger proportion of buildings with a peak consumption hour in the late morning or early afternoon, earlier than the typical afternoon peak seen in other cities. Beer Sheva and Netanya also stand out for having a higher minimum consumption hour, with the daily trough occurring later in the morning rather than in the pre-dawn hours. A plausible explanation is that these cities contain a larger share of buildings with continuous daytime occupancy, which sustain consumption throughout the day.

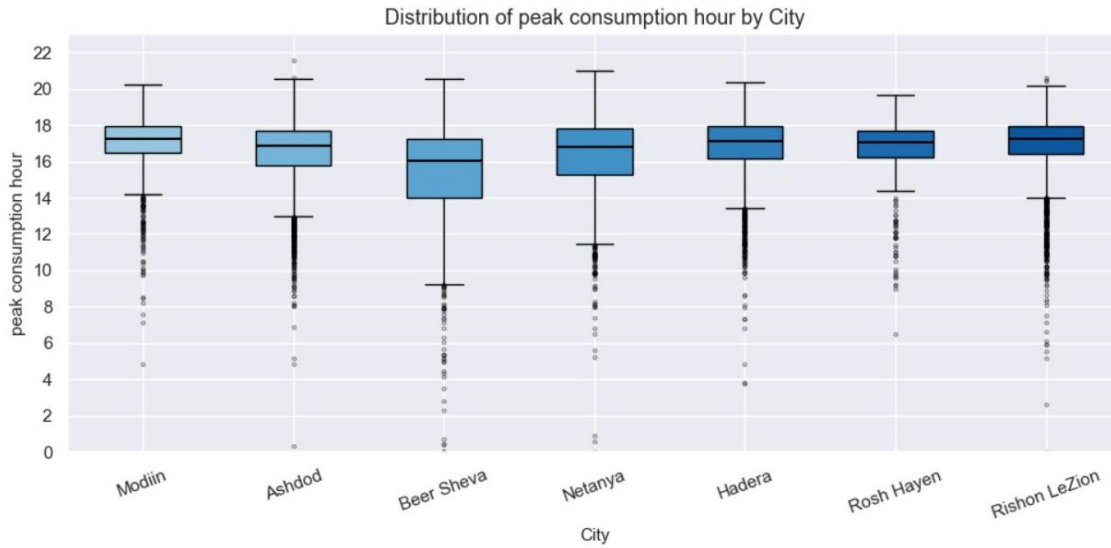


Figure 17: Distribution of peak consumption hour by city.

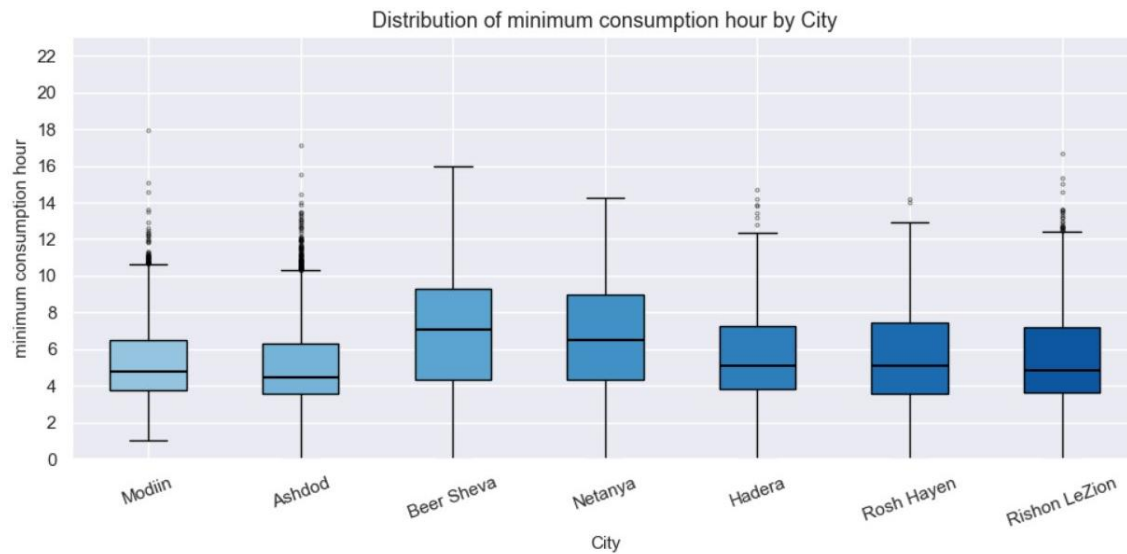


Figure 18: Distribution of minimum consumption hour by city.

To further characterise inter-city differences in consumption dynamics, we apply the Prophet (Taylor & Letham, 2018) decomposition to the city-level mean consumption series. Prophet is an additive time series model that represents observed consumption as the sum of a long-term trend component, weekly seasonality, daily seasonality, and a zero-mean

noise term. This decomposition allows direct comparison of each component's city-level contribution across cities.

Figure 19, Figure 20, and Figure 21 present a Prophet decomposition of the city-level mean consumption series into 3 components: Trend, Weekly seasonality, and Daily seasonality. Figure 19 presents the trend component and confirms that all cities follow the same annual arc, peaking in summer and troughing in winter, with Rosh Ha'ayin consistently elevated throughout the year. Figure 20 (the weekly seasonality component) shows that while most cities exhibit a positive Friday–Saturday contribution, Beer Sheva, Netanya, and Rosh Ha'ayin display a negative weekend effect (Beer Sheva most strongly), reinforcing the pattern identified in Figure 15. Figure 21 (the daily seasonality component) most clearly distinguishes Beer Sheva from the rest: it shows the largest morning peak (centred around 11:00–12:00) and the steepest evening decline of any city, with a negative contribution already from 16:00 onward. Rosh Ha'ayin also displays an elevated morning peak, while Modiin's daily profile is more evenly spread across the afternoon hours.

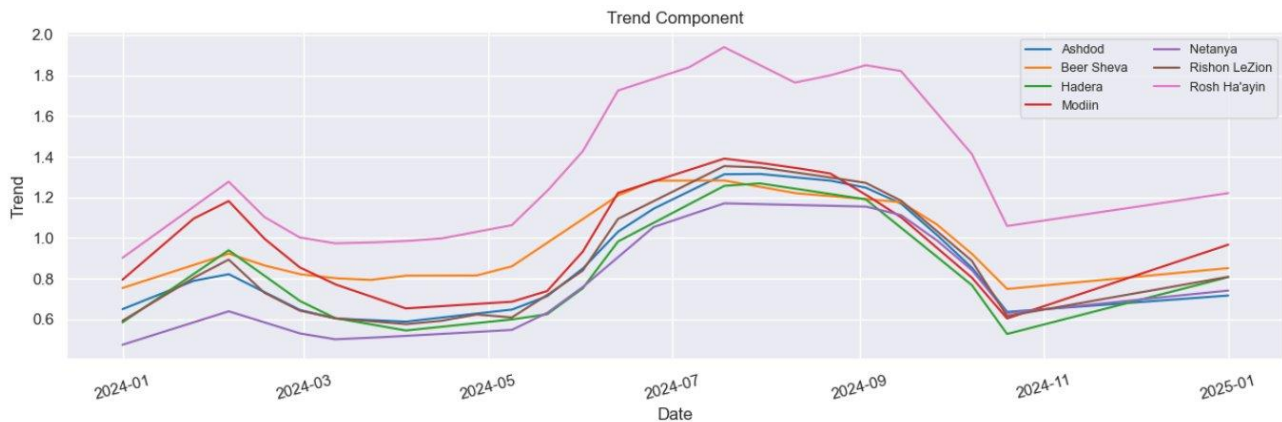


Figure 19: Prophet decomposition by city (2024) – Trend component. We did not run a similar analysis on the 2023 dataset because of significant gaps in January and February 2023.

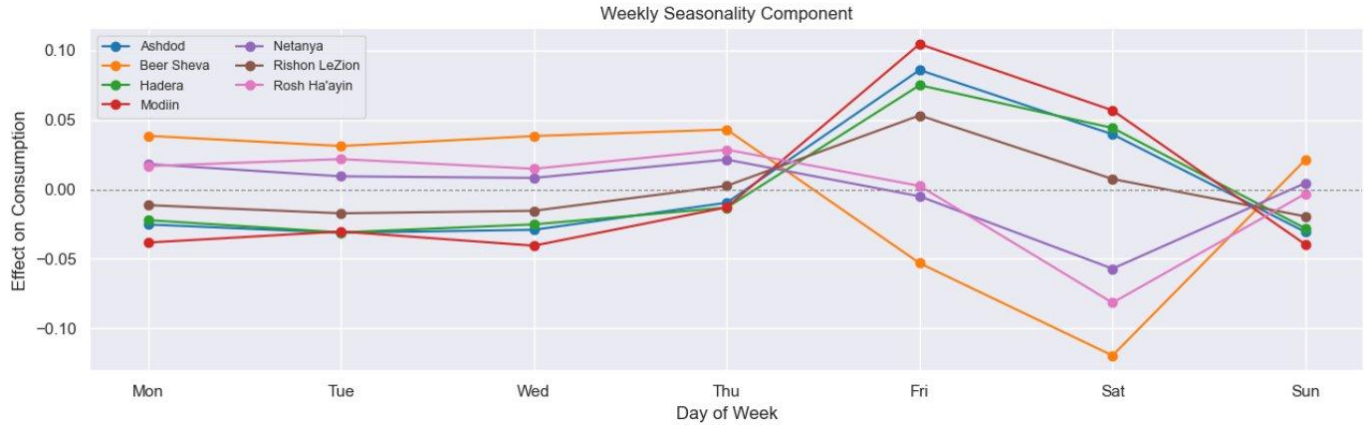


Figure 20: Prophet decomposition by city (2024) – Weekly seasonality component. We did not run a similar analysis on the 2023 dataset because of significant gaps in January and February 2023.

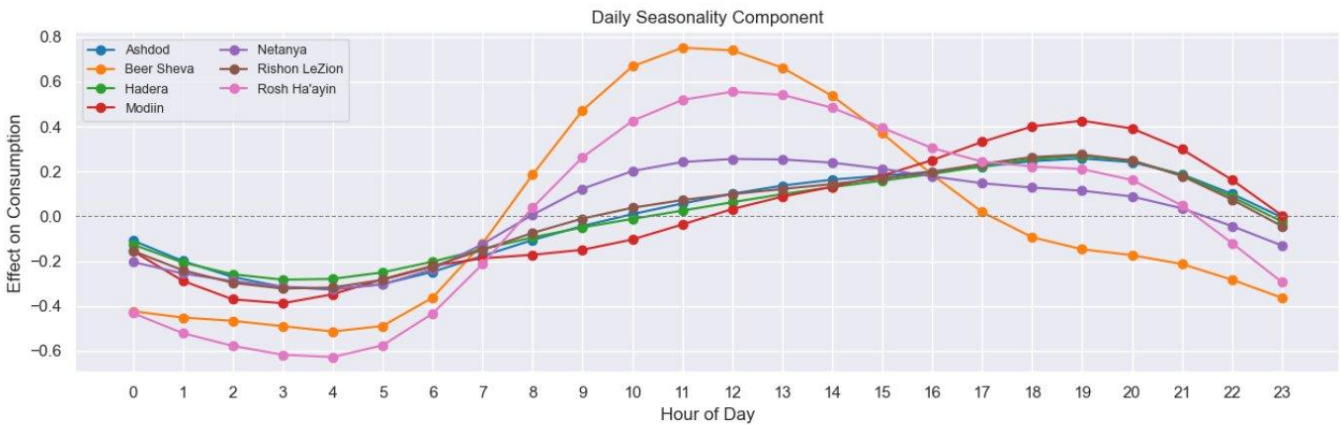


Figure 21: Prophet decomposition by city (2024) – Daily seasonality component. We did not run a similar analysis on the 2023 dataset because of significant gaps in January and February 2023.

2.3.4 Apartment floor area and electricity consumption

Beyond temporal and weather-related patterns, we examined the relationship between apartment size and average annual electricity consumption per floor area. In a previous study on electricity consumption patterns in Israeli apartments, based on a representative sample of 822 households, electricity consumption per occupant was higher in larger apartments (Bugin et al., 2022). The authors hypothesised that larger apartments were occupied by higher-income families, who tend to use more energy for air conditioning and electric heating, along with other electrical appliances. The authors also found a weaker

opposite relation between consumption per m^2 and an apartment's total floor area: in smaller apartments, consumption normalised by floor area tended to be higher.

Because our dataset covered many more meters, we tested the previous study's finding by examining the relationship between apartment size and annual hourly consumption per m^2 . Because we could not retrieve personal data on household occupancy and actual apartment size to protect privacy, we assumed that each meter in a building represents an apartment and that each apartment is identical in size. We then divided the total building floor area, calculated according to the method described in Section 2.1.3 by the number of meters in that building, to generate an average apartment size for that building. Similarly, we calculated the average hourly consumption per meter for 2024 and normalised it separately for each building by the calculated average apartment size. We chose 2024 for this analysis because we had full annual coverage of meter readings for a substantial number of buildings, unlike in 2023.

For the sake of clarity and to avoid outliers, the generated chart (Figure 22) presents only apartments below 500 m^2 , hourly consumption below 100 Wh/m^2 , and buildings with at least 8,400 hours of consumption data in 2024. Overall, this analysis covered 206,671 meters across 12,709 buildings. The chart clearly shows that as apartments grow larger in floor area, consumption normalised by floor area decreases. This finding is consistent with the findings of the previous study mentioned above (Bugin et al., 2022), although they seem more pronounced.

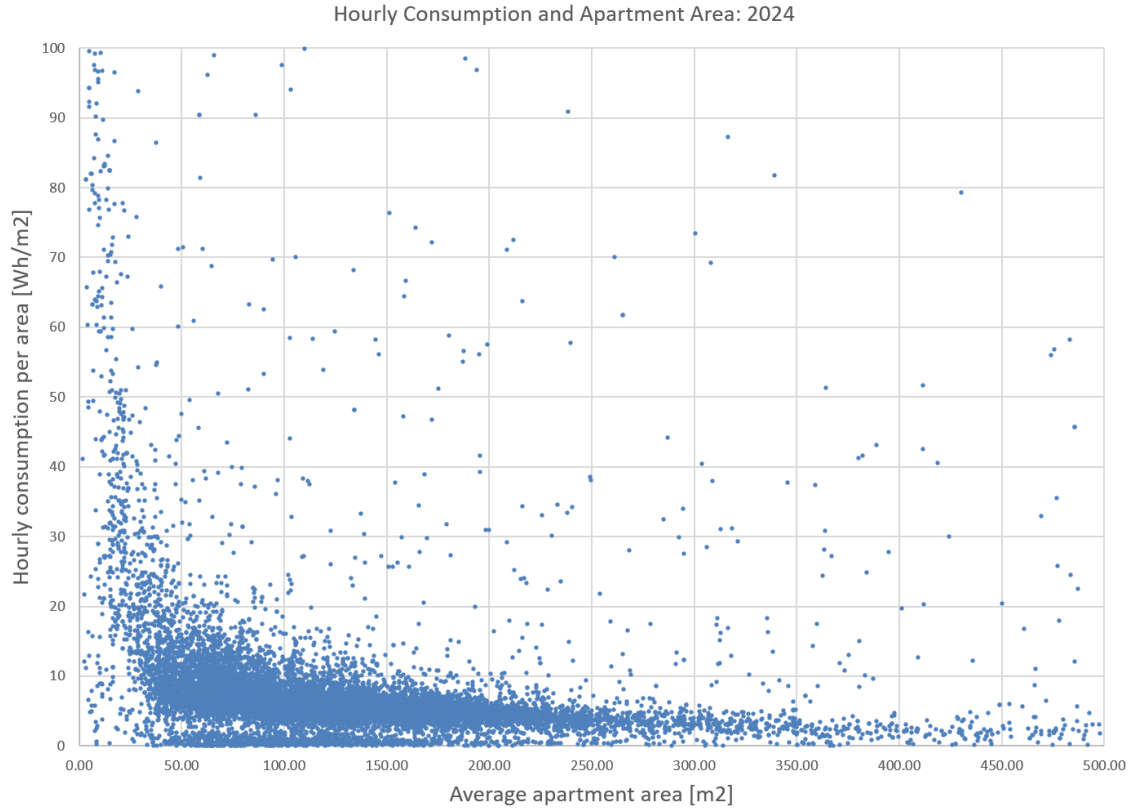


Figure 22: The relation between consumption per area and apartment floor area, generated from the 2024 dataset (n=206,671 meters, 12,709 buildings).

2.3.5 Feature extraction from consumption time series – tsfresh

This analysis aimed to construct a feature vector that characterises each building based solely on its electricity consumption time series, enabling buildings to be compared and grouped by consumption behaviour. To do this, we extracted a rich statistical feature space from each building's normalised consumption time series (hourly consumption divided by the number of registered meters) using the tsfresh package (Time Series Feature extraction based on Scalable Hypothesis tests) (Christ et al., 2018). In total, we extracted 782 features per building. We analysed a subset of 5,899 buildings with continuous data across the full analysis period. We ran the feature extraction process twice: once for 2023 time-series consumption and once for 2024 time-series consumption. For 2023, we extracted features only from March–December because of the large number of outliers and missing values in January–February. For 2024, we used the full year. After feature extraction, NaN values and

infinite values were removed, and feature vectors were standardised using StandardScaler (zero mean, unit variance). Figure 23 and Figure 24 show the t-SNE projection of the tsfresh feature space, coloured by city.

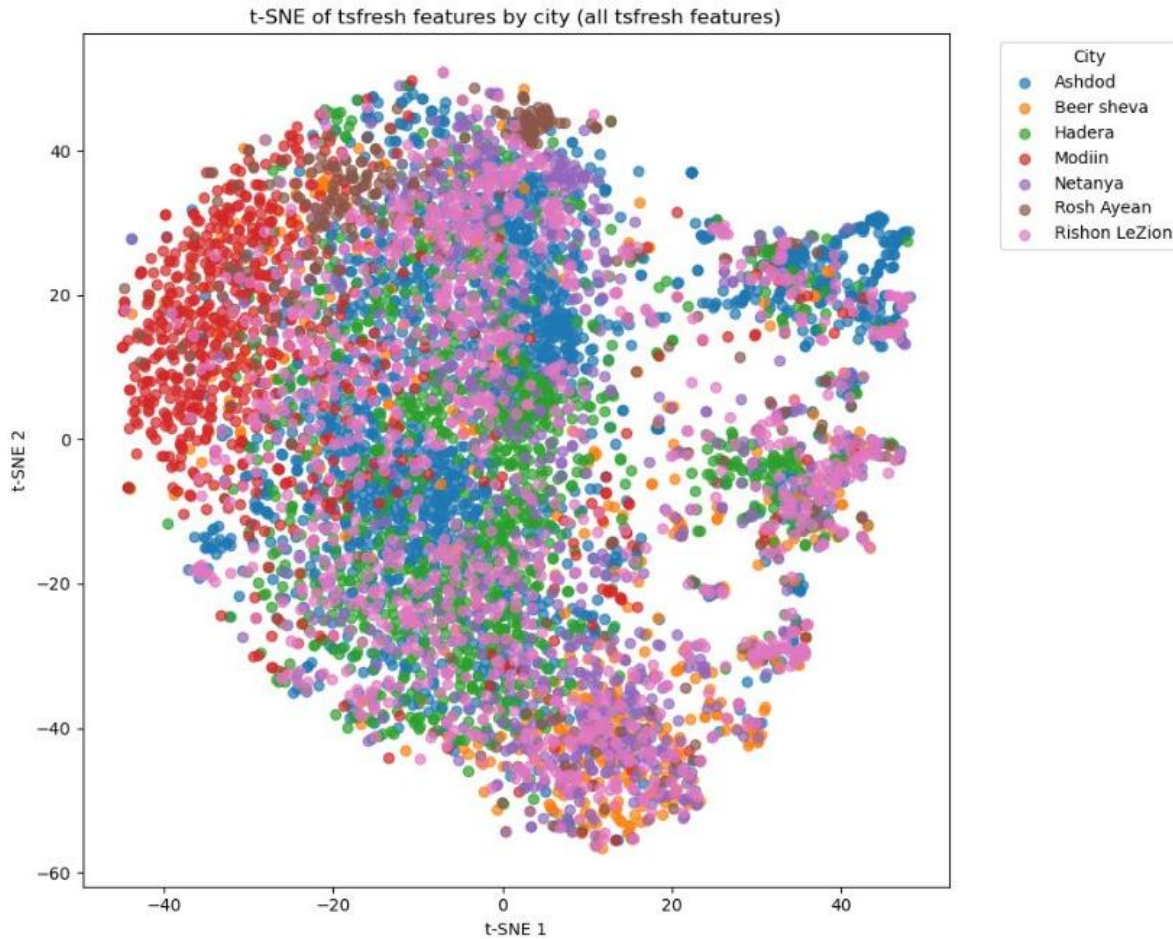


Figure 23: t-SNE projection of the tsfresh feature space coloured by city, 2023.

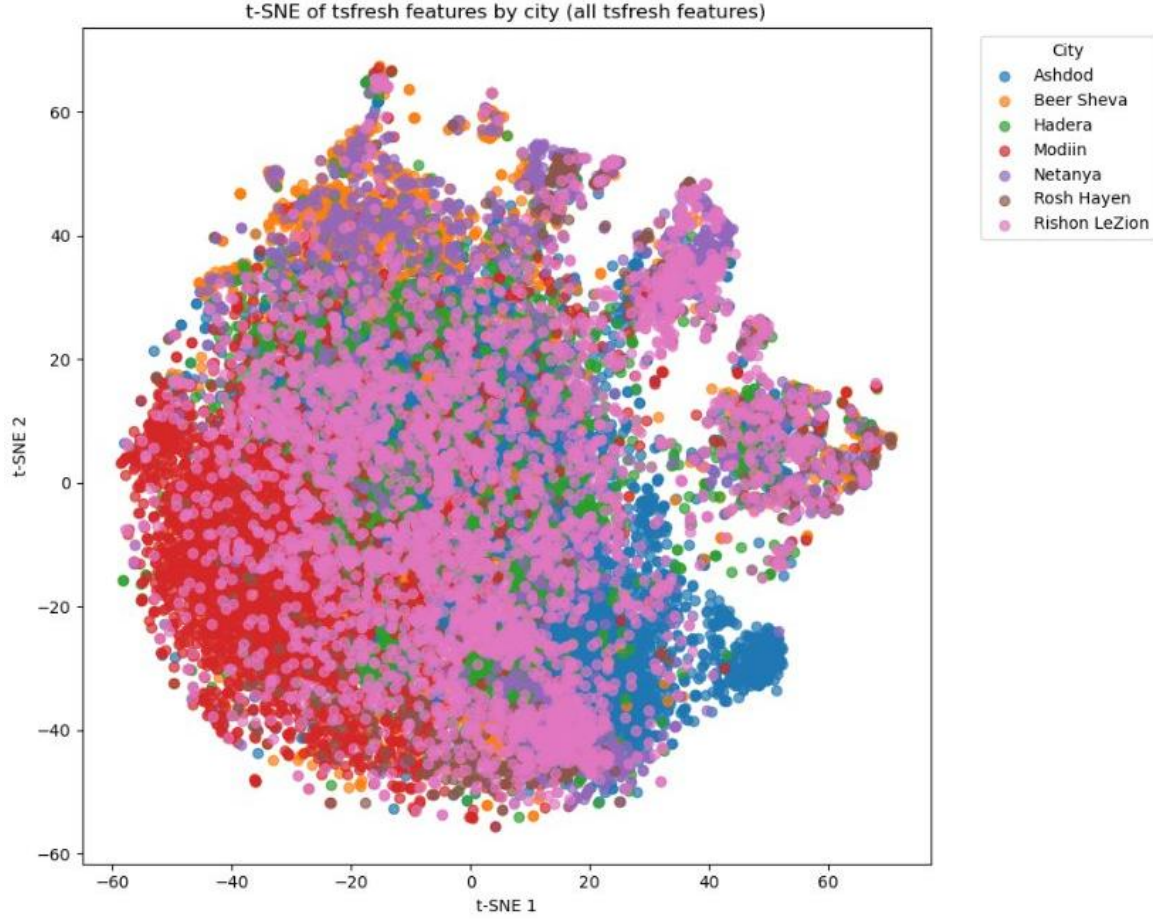


Figure 24: t-SNE projection of the tsfresh feature space coloured by city, 2024.

The separability of the feature space was assessed visually via 3 metrics, as follows:

- **Mean distance to centroid** measures the average Euclidean distance between each building's feature vector and its city's centroid in the standardised feature space. For city C_i with centroid μ_i , the mean distance is:

$$d_i = \frac{1}{|C_i|} \sum \|f_b - \mu_i\|_{b \in C_i}$$

where f_b is the feature vector of building b and $|C_i|$ is the number of buildings in a city C_i . Lower values indicate tighter, more cohesive within-city clusters. Table 5 reports the mean distance to centroid in the standardised feature space, by city and year.

- **kNN@10** same-city fraction is the proportion of a building's 10 nearest neighbours (in feature space) that belong to the same city. For building b :

$$p_b = \frac{1}{k} \sum_{j=1}^k 1[\text{city}(n_j(b)) = \text{city}(b)]$$

where $n_j(b)$ denotes the j -th nearest neighbour of b in feature space.

The balanced score normalises this fraction by the city's prevalence in the dataset, correcting for class-size differences. Table 6 reports the kNN@10 same-city fraction and balanced score, by city and year.

- **The Silhouette score** for building i is defined as:

$$\frac{b(i) - a(i)}{\max(b(i), a(i))}$$

where $a(i)$ is the mean distance from i to all other buildings in the same city, and $b(i)$ is the mean distance from i to all buildings in the nearest other city. The score ranges from -1 to 1 ; positive values indicate that a building is on average closer to its own city than to any other, while negative values indicate the opposite. Table 7 reports the mean silhouette per city and year.

Table 5, Table 6, and Table 7 show the results of this evaluation process. Taken together, the separability metrics reveal a consistent picture across both years. Ashdod, Modiin, and Hadera show the tightest within-city clusters (lowest mean distance to centroid) and the highest same-city kNN fractions, indicating that buildings from these cities share more distinctive and cohesive consumption patterns. Beer Sheva is consistently the most dispersed city, with the largest mean centroid distance and the lowest kNN same-city fraction in both years, suggesting high within-city heterogeneity or strong overlap with other cities in feature space. At the individual city level, Hadera and Modiin show positive silhouette scores in 2023, confirming their relative distinctiveness, while Beer Sheva has the most negative score. These findings are consistent with the city classification performance reported in Section 2.3.6.

Table 5: Mean distance to centroid in the standardised feature space, by city and year

City	Year	Mean Distance to Centroid
Ashdod	2023	16.99
Ashdod	2024	14.73
Beer Sheva	2023	24.97
Beer Sheva	2024	27.92
Hadera	2023	14.98
Hadera	2024	15.10
Modiin	2023	16.19
Modiin	2024	15.19
Netanya	2023	20.91
Netanya	2024	22.13
Rishon LeZion	2023	20.19
Rishon LeZion	2024	15.40
Rosh Ha'ayin	2023	16.89
Rosh Ha'ayin	2024	22.37

Table 6: kNN@10 same-city fraction and balanced score, by city and year

City	Year	kNN@10 Same-City Fraction	Balanced Score
Ashdod	2023	0.562	0.383
Ashdod	2024	0.553	0.391
Beer Sheva	2023	0.197	0.136
Beer Sheva	2024	0.303	0.235
Hadera	2023	0.433	0.313
Hadera	2024	0.416	0.326
Modiin	2023	0.331	0.250
Modiin	2024	0.528	0.433
Netanya	2023	0.252	0.152
Netanya	2024	0.309	0.242
Rishon LeZion	2023	0.392	0.251
Rishon LeZion	2024	0.557	0.426
Rosh Ha'ayin	2023	0.349	0.316
Rosh Ha'ayin	2024	0.255	0.233

Table 7: Mean silhouette score per city, by year

City	Year	Mean Silhouette Score
Ashdod	2023	-0.060
Ashdod	2024	0.027
Beer Sheva	2023	-0.198
Beer Sheva	2024	-0.255
Hadera	2023	0.070
Hadera	2024	-0.003
Modiin	2023	0.018
Modiin	2024	0.016
Netanya	2023	-0.156
Netanya	2024	-0.158
Rishon LeZion	2023	-0.155
Rishon LeZion	2024	-0.026
Rosh Ha'ayin	2023	-0.031
Rosh Ha'ayin	2024	-0.199

2.3.6 Building electricity consumption classification by city

A key question is whether a building's electricity-consumption fingerprint, encoded in the `tsfresh` feature vector, is distinctive enough to identify its city of origin without any geographic information. To test this, we framed the problem as a 7-class classification task predicting each building's city of origin. An SVM classifier with an RBF kernel (Hyperparameters: $C = 30$, $\gamma = 0.001$, `class_weight='balanced'`) was trained using 5-fold Cross-Validation, with features standardised within each fold.

We measured performance using Macro F1 and Weighted F1. Macro F1 is the unweighted average of the per-class F1 scores, treating all classes equally regardless of size; it is sensitive to performance on small classes. Weighted F1 is the class-size-weighted average of per-class F1 scores, giving more weight to larger classes and reflecting overall accuracy on the full dataset. Table 8 reports the classification-by-city task results. Figure 25 and Figure 26 present the confusion matrices.

Table 8: Classification by city results

Year	Macro Precision	Macro Recall	Macro F1	Weighted F1
2023 (March–December)	0.725	0.705	0.713	0.745
2024	0.773	0.762	0.766	0.794

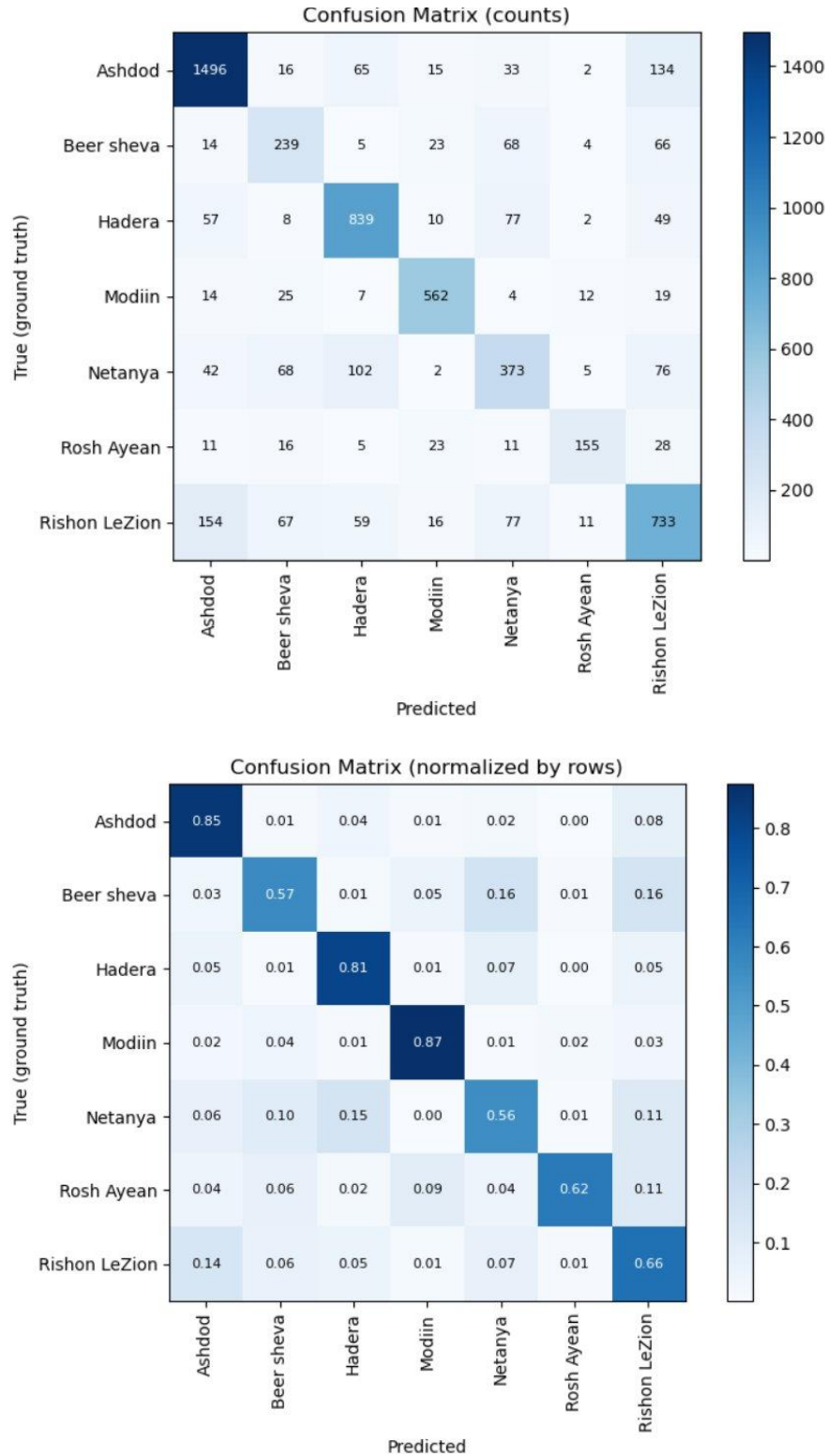


Figure 25: Confusion matrices for city classification, 2023. Top: raw counts. Bottom: row-normalised (recall per city).

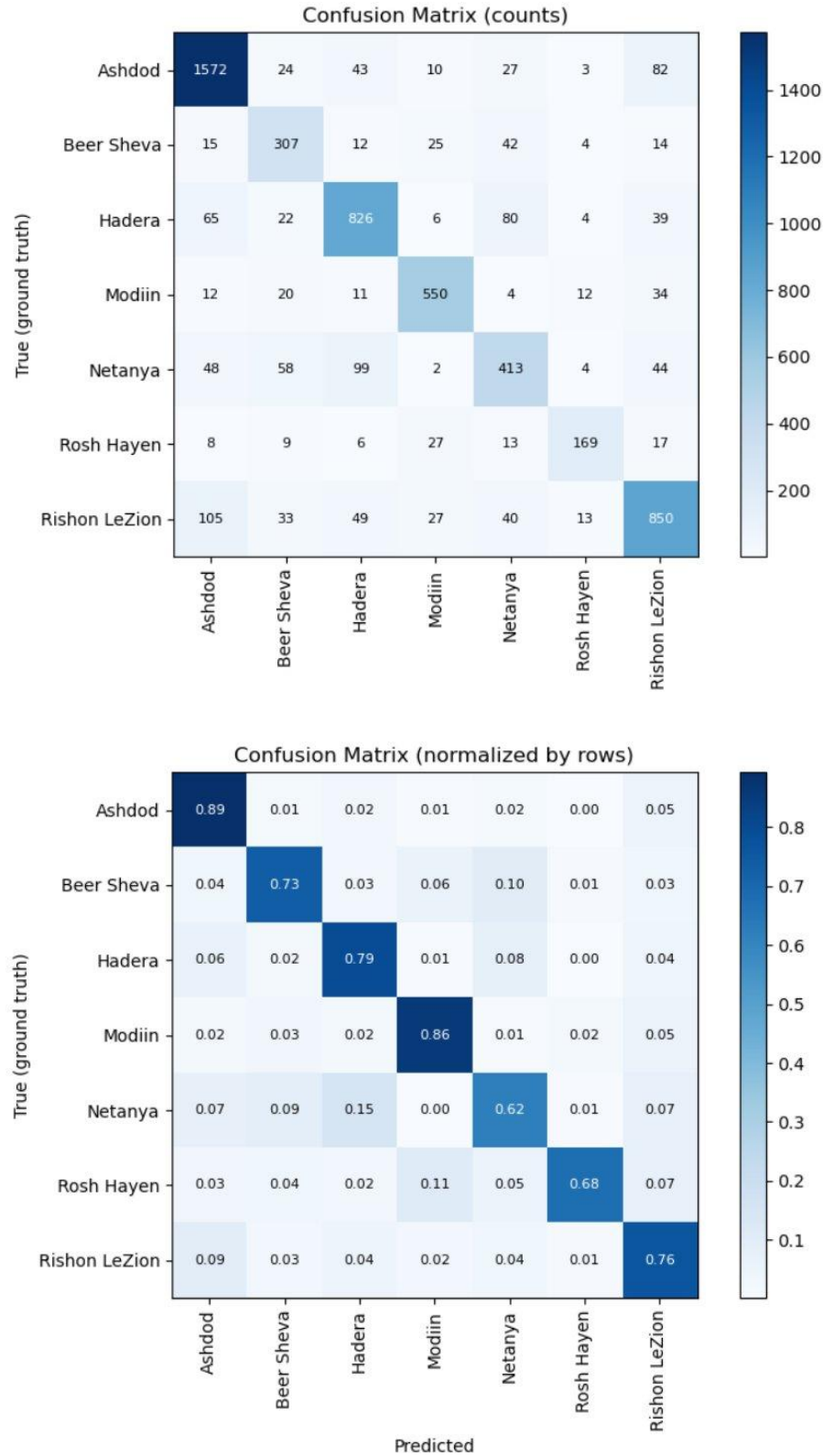


Figure 26: Confusion matrices for city classification, 2024. Top: raw counts. Bottom: row-normalised (recall per city).

The results show strong classification performance, with the model performing better in 2024 than in 2023 (weighted F1 score of 0.745 in 2023 and 0.794 in 2024). The confusion matrices reveal consistent patterns across both years. Ashdod and Modiin are the best-classified cities, with row-normalised recall of 0.85 and 0.87 in 2023 and 0.89 and 0.86 in 2024, reflecting their distinctive consumption fingerprints. Hadera is also well identified (recall 0.81 in 2023, 0.79 in 2024). These results align with the measures calculated in section 2.3.5, which showed these cities are the most separable. In contrast, Rishon LeZion shows lower recall (0.66 in 2023, 0.76 in 2024), with a notable proportion of its buildings misclassified as Ashdod, suggesting overlap in consumption patterns between these 2 cities. Beer Sheva improves substantially from 2023 to 2024 (0.57 to 0.73). Netanya and Rosh Ha'ayin show moderate recall in both years (0.56–0.68). Overall, classification accuracy was higher for the 2024 dataset.

2.3.7 Socioeconomic cluster classification

Beyond geographic origin classification, we also asked whether a building's consumption fingerprint encodes information about the socioeconomic character of its neighbourhood. Specifically, we used ordinal regression to predict the socioeconomic cluster of the statistical zone (stat area) in which each building is located, ranging from 2 to 8 (There is also a socioeconomic cluster 9, but we removed these buildings from this model because less than 1% of the buildings in the dataset belong to it), based solely on its tsfresh feature vector. Because the target variable is ordinal, we used ordinal regression. In addition to the classification evaluation measures Macro F1 and Weighted F1, we also evaluate the model using regression measures, mean absolute error (MAE), mean squared error (MSE), and R^2 , which penalise predictions based on their distance from the true cluster. Figure 9 reports the Socioeconomic Cluster Classification task's results. Figure 27 and Figure 28 present the confusion matrices.

Table 9: Socioeconomic cluster classification results

Year	Macro F1	Weighted F1	MAE	MSE	R²
2023 (March–December)	0.469	0.507	0.938	2.405	0.189
2024	0.501	0.542	0.822	1.951	0.334

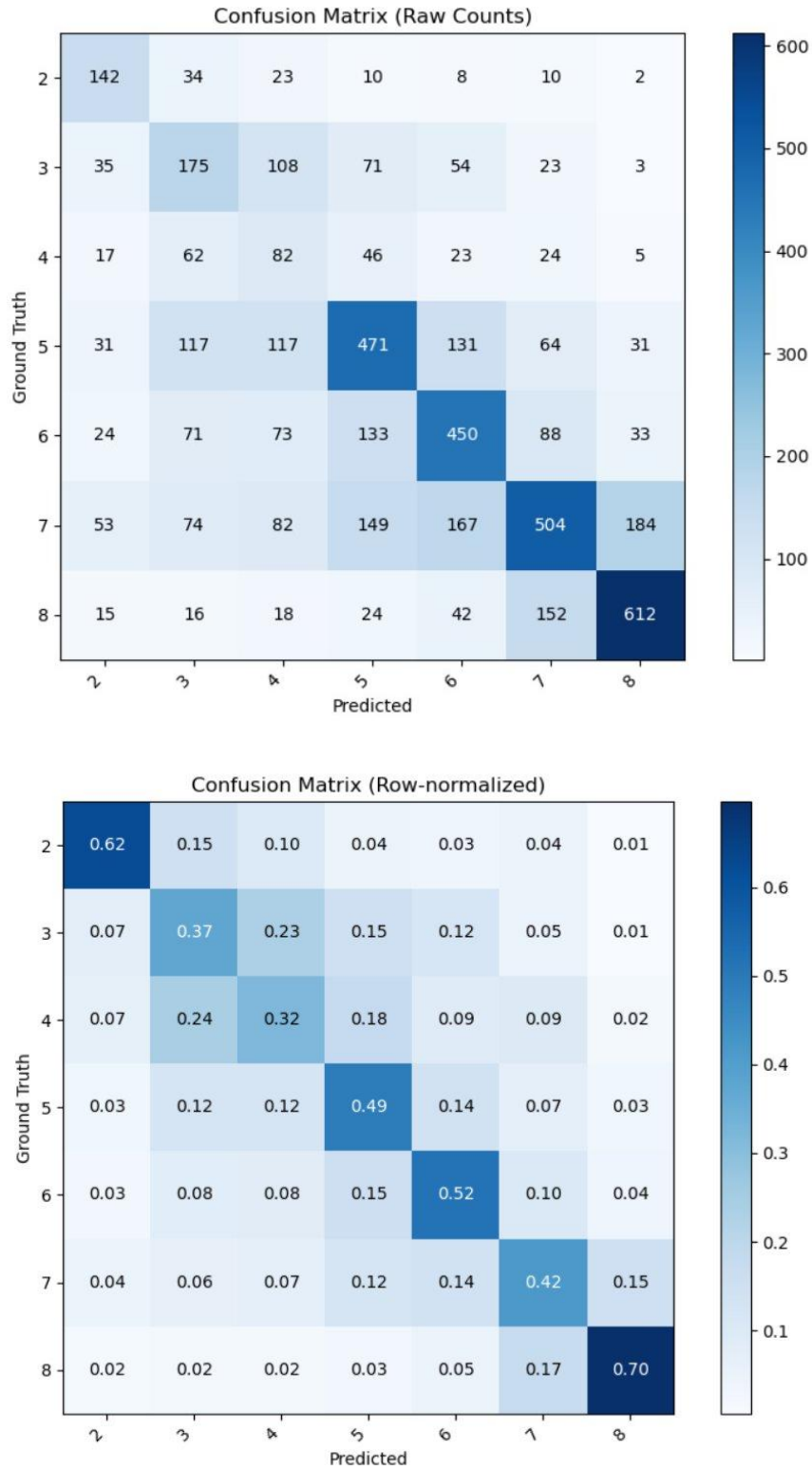


Figure 27: Confusion matrices for socioeconomic cluster classification, 2023. Top: raw counts. Bottom: row-normalised (recall per cluster).

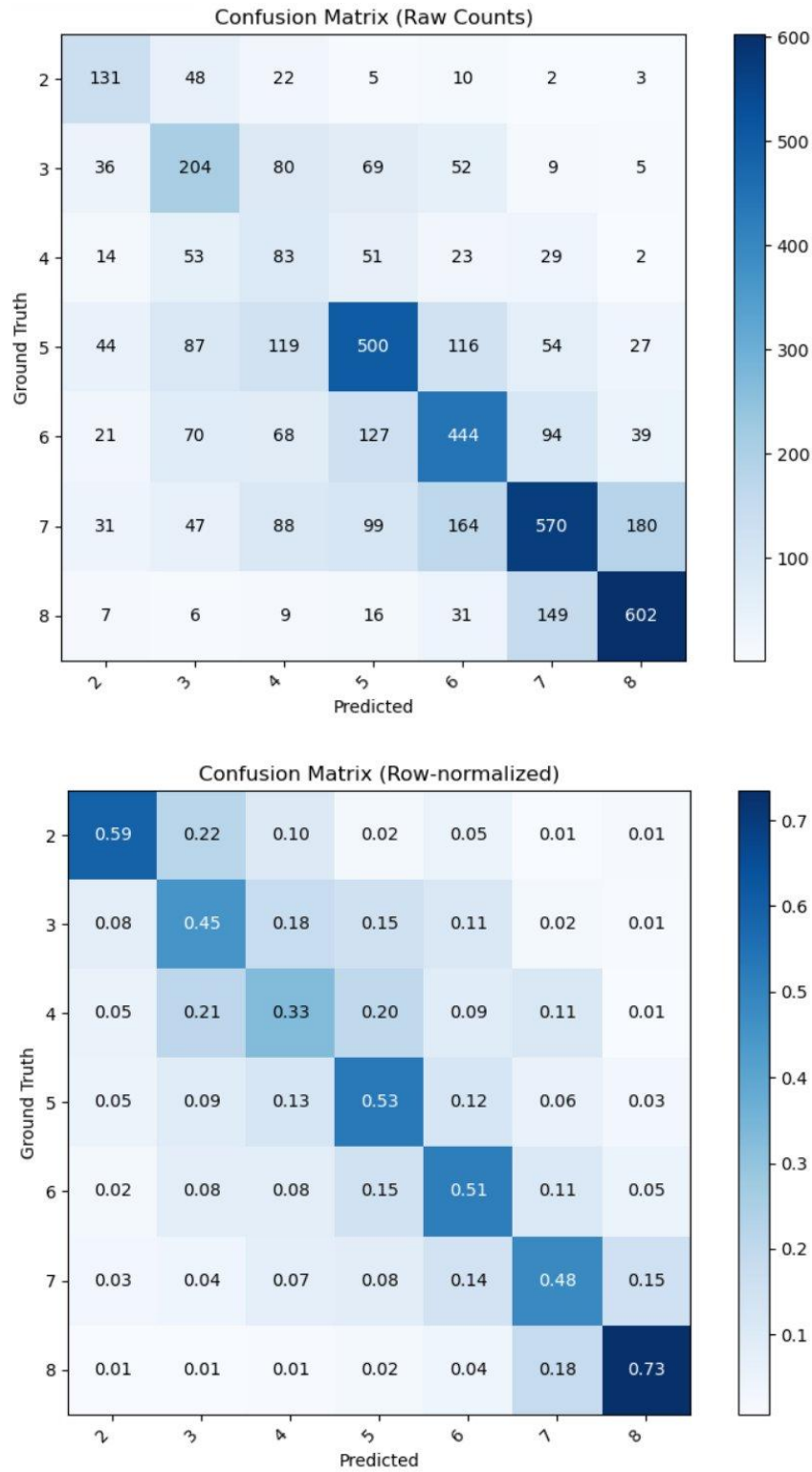


Figure 28: Confusion matrices for socioeconomic cluster classification, 2024. Top: raw counts. Bottom: row-normalised (recall per cluster).

The socioeconomic confusion matrices show a clear pattern: the extreme clusters are recalled substantially better than those in the middle. Cluster 8 (highest socioeconomic status) achieves recall of 0.70 in 2023 and 0.73 in 2024, and cluster 2 (lowest) achieves 0.62 and 0.59, respectively. One likely explanation for the high precision in cluster 8 is that the entire Modiin building dataset belongs to this cluster, and, as shown in the inter-city analysis, Modiin displays distinctively homogeneous consumption patterns that differ from other cities in the dataset. For cluster 2, a plausible reason for its relatively high recall is the presence of ultra-Orthodox Jewish neighbourhoods within this socioeconomic cluster, which typically exhibit consumption patterns that differ substantially from the general population due to distinct lifestyle and occupancy patterns. The middle clusters (3-7) are harder to classify, with recall typically ranging from 0.32 to 0.53, and misclassifications tend to occur between adjacent clusters, consistent with the task's ordinal nature. As in the city classification task, performance was better on the 2024 dataset than on the 2023 dataset, as reflected in higher Macro F1 (0.469 $\rightarrow$ 0.501), lower MAE (0.938 $\rightarrow$ 0.822), and higher R^2 (0.189 $\rightarrow$ 0.334).

2.3.8 Classifying types of construction-consumption nexus

This cluster-based classification aimed to identify distinct types of the “construction-consumption nexus”, defined as the combined effect of construction technology and occupant consumption behaviour on a building’s energy consumption levels. The limited availability of large, open datasets on construction technology, building age, or occupant socioeconomics and demographics at the housing-unit scale makes it practically impossible to assess the effects of these characteristics separately on electricity use. Therefore, to analyse recurring trends across the dataset, we use the building (not the apartment or its occupants) as the core entity for classification, knowing that it “fuses” the varied effects of the construction features of all its apartments and the consumption traits of its anonymous occupants.

To classify buildings' statistical consumption fingerprints, we used the same standardised tsfresh feature space described in Section 2.3.5, computed over the consumption series from March 2023 to December 2024, for the same subset of 5,899

buildings with continuous data across this period. We applied K-Means clustering to the tsfresh feature space to identify natural groupings of buildings based on their consumption patterns. To manage the high dimensionality (782 features), we first standardised all features with StandardScaler, then applied PCA to reduce the representation to 50 dimensions. K-Means was run with $k = 18$. We then merged clusters with fewer than 50 buildings into their nearest-neighbour cluster in the 50-dimensional PCA space, based on centroid distance, to avoid a small number of outlier buildings forming their own cluster. This merging process yielded 7 substantive, interpretable clusters. Results were visualised by projecting the feature space onto the t-SNE coordinates computed in Section 2.3.5. Figure 29 shows the projection across all 5,899 buildings, and Figure 30 shows the same projection separately for each of the 7 cities. Figures 31-37 show the geographic distribution of the same clusters by city.

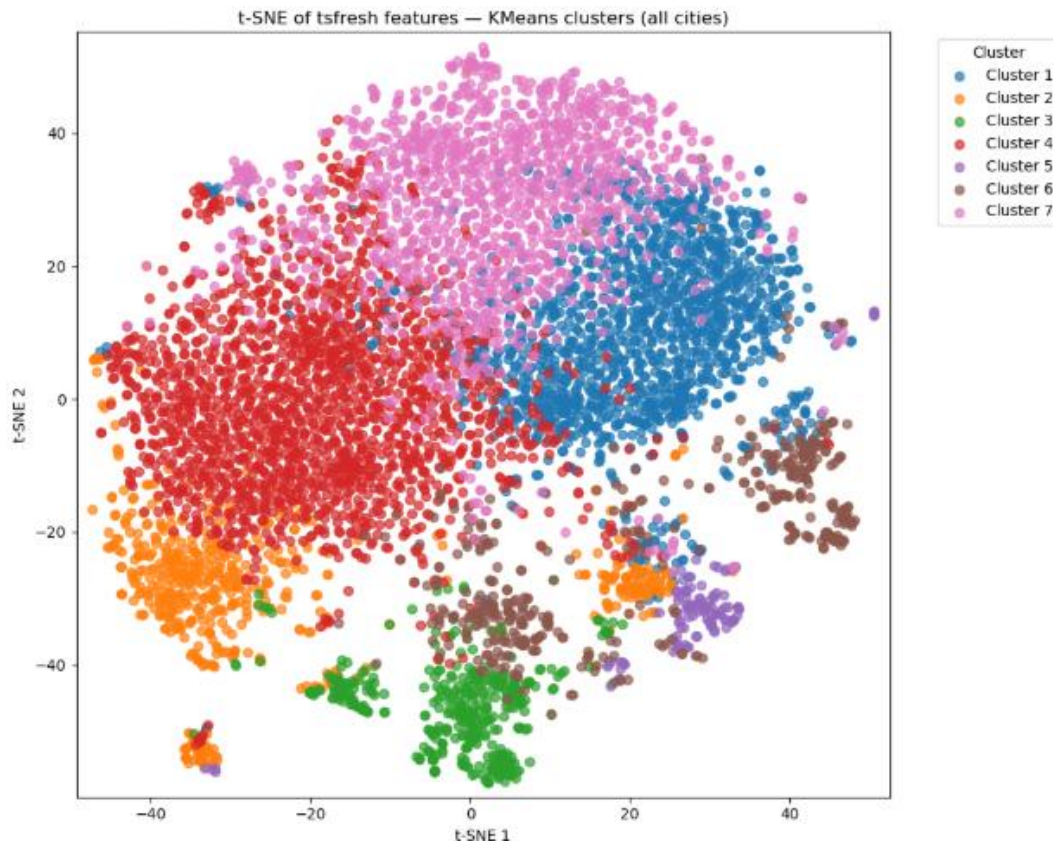


Figure 29: t-SNE projection of the tsfresh feature space, coloured by K-Means cluster assignment, across all buildings.

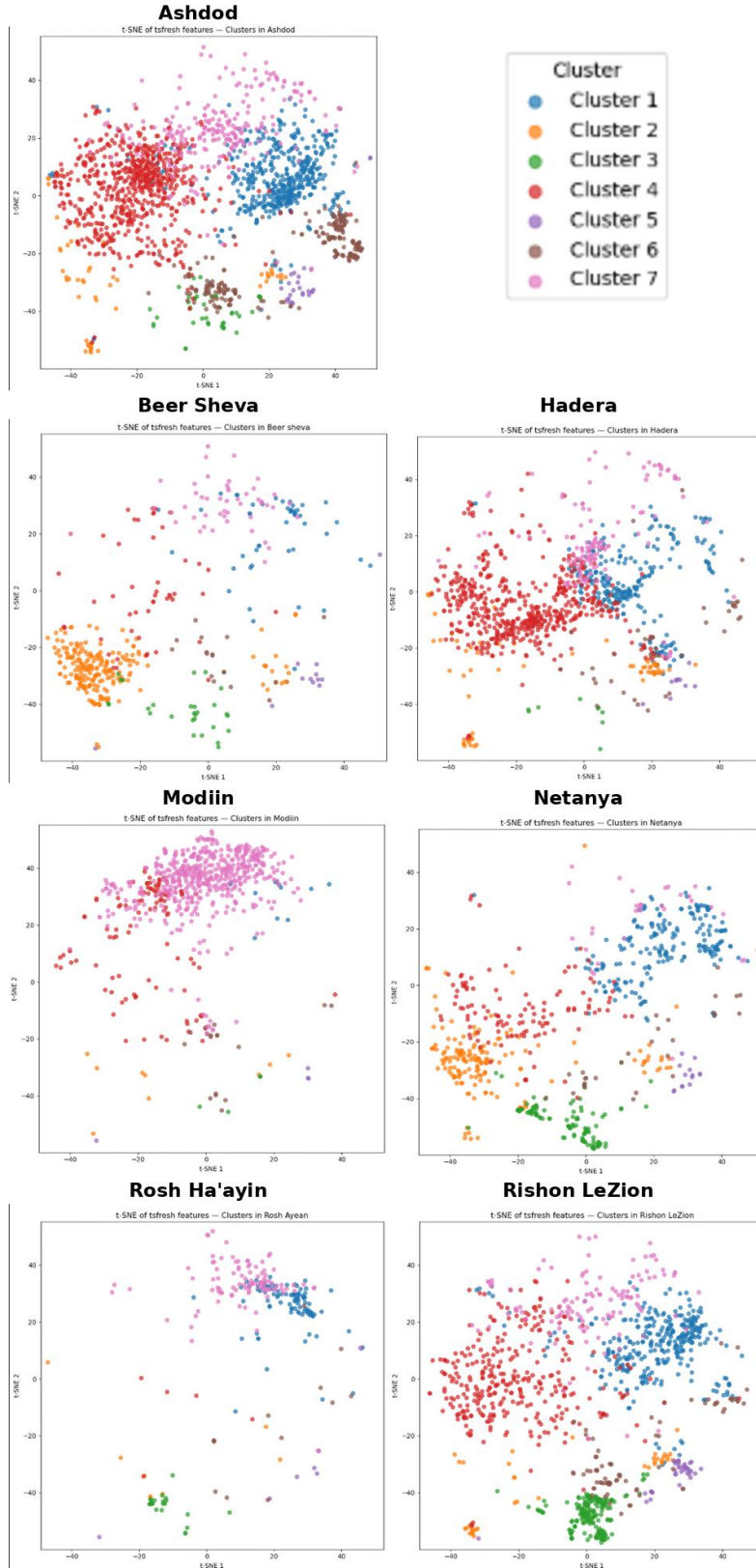


Figure 30: t-SNE projection of the tsfresh feature space, coloured by K-Means cluster assignment, shown separately for each city.

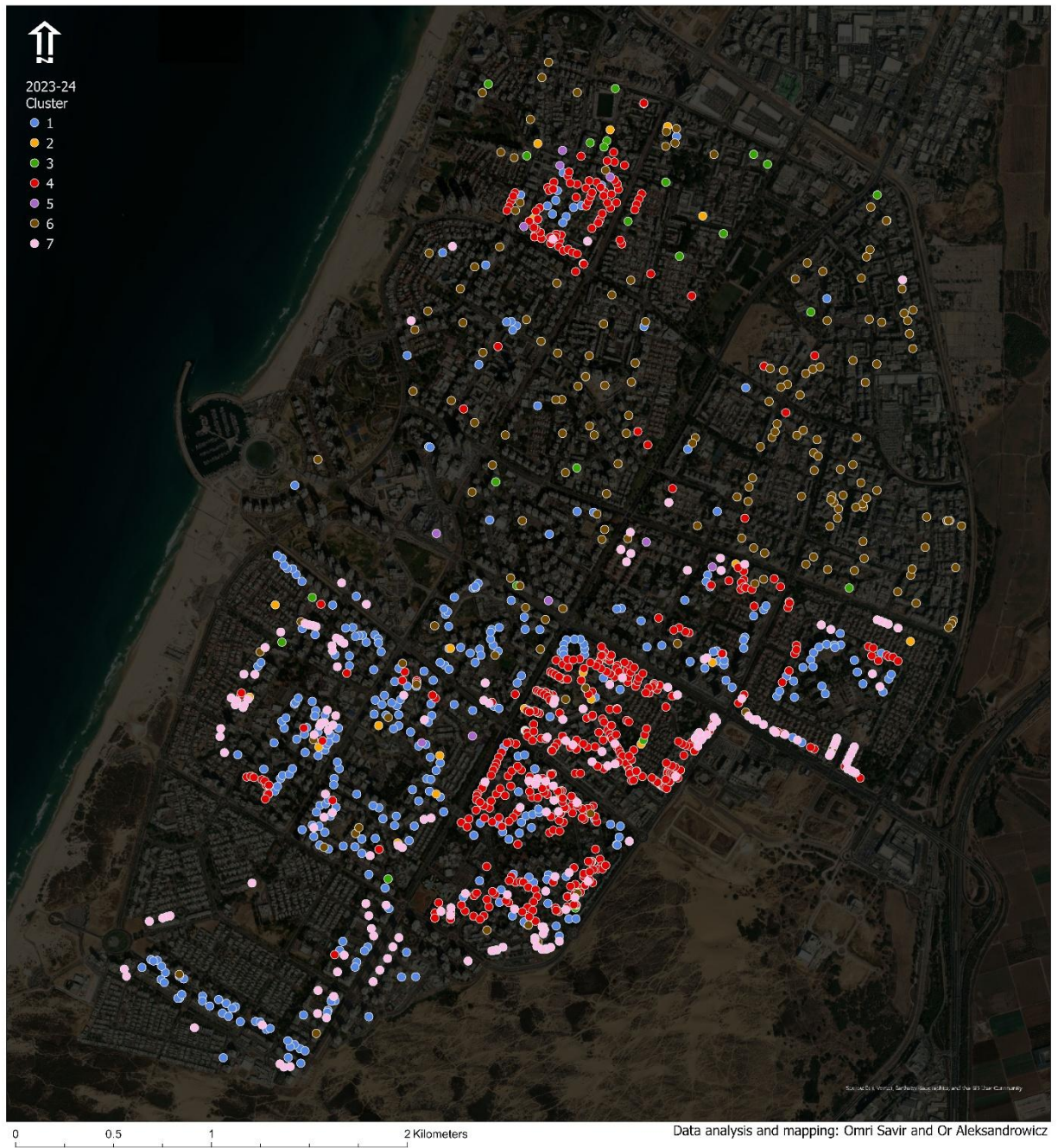


Figure 31: Geographic distribution of the K-Means clustering of the dataset buildings in Ashdod.

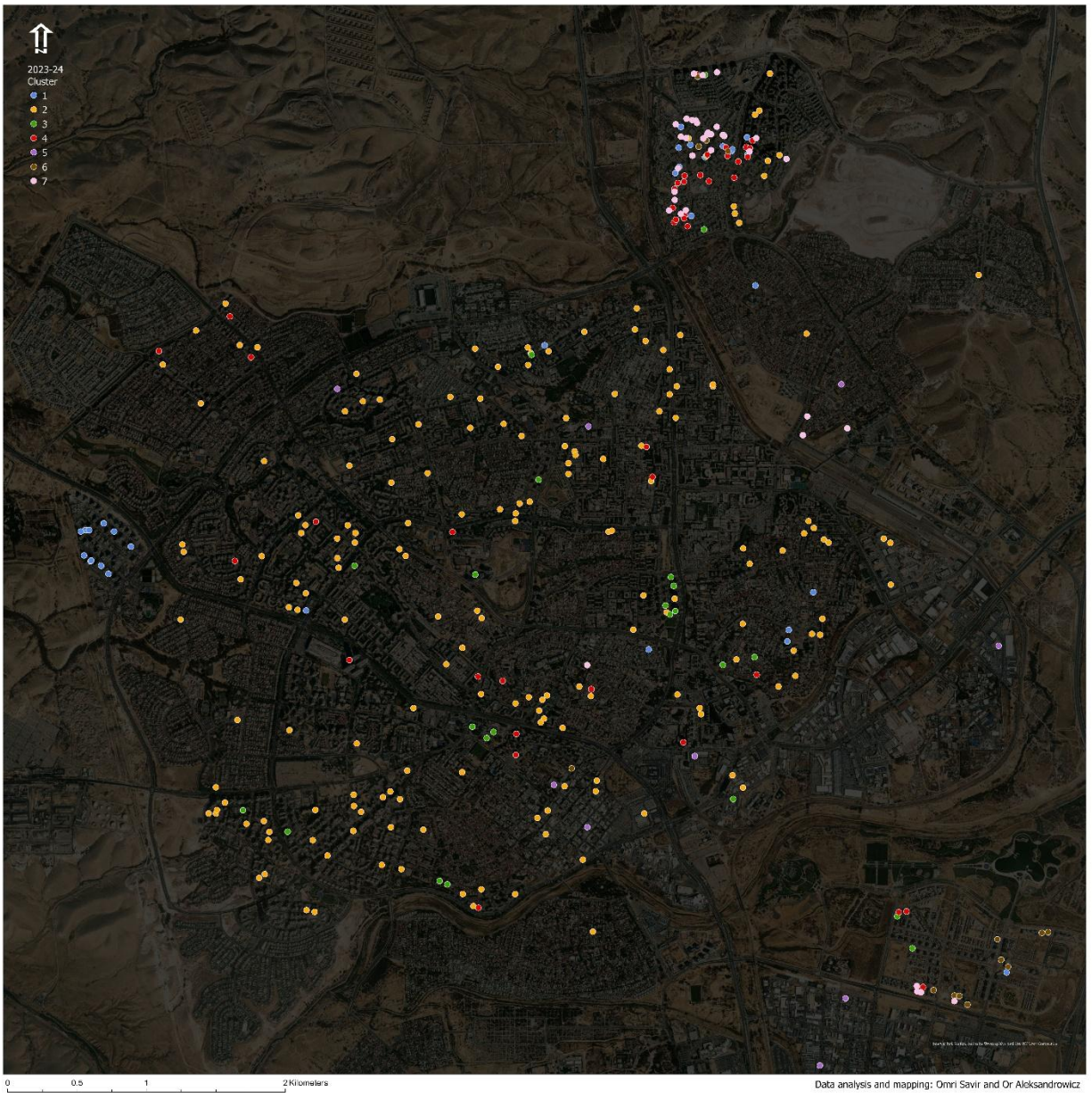


Figure 32: Geographic distribution of the K-Means clustering of the dataset buildings in Beer Sheva.

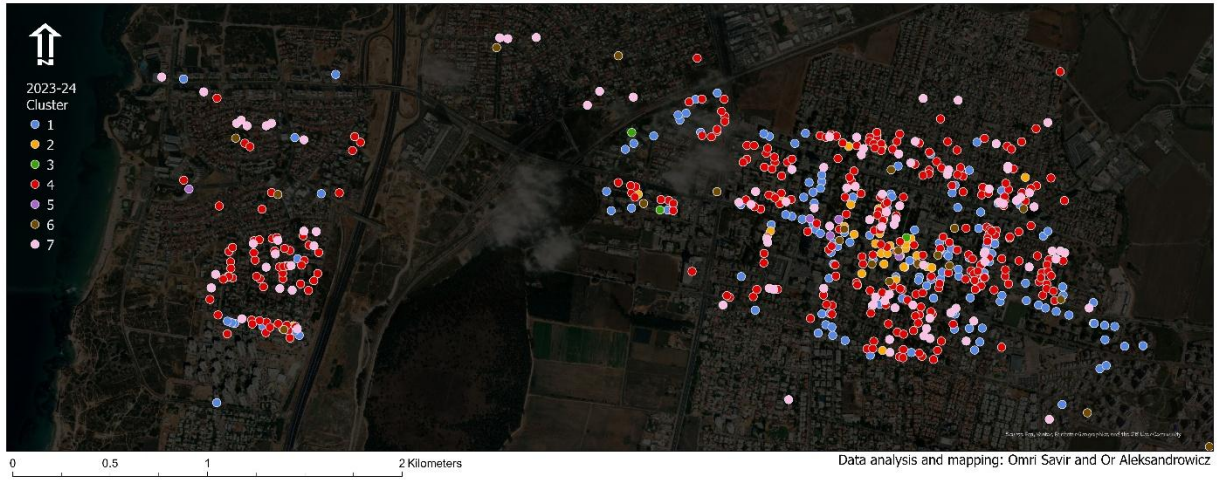


Figure 33: Geographic distribution of the K-Means clustering of the dataset buildings in Hadera.

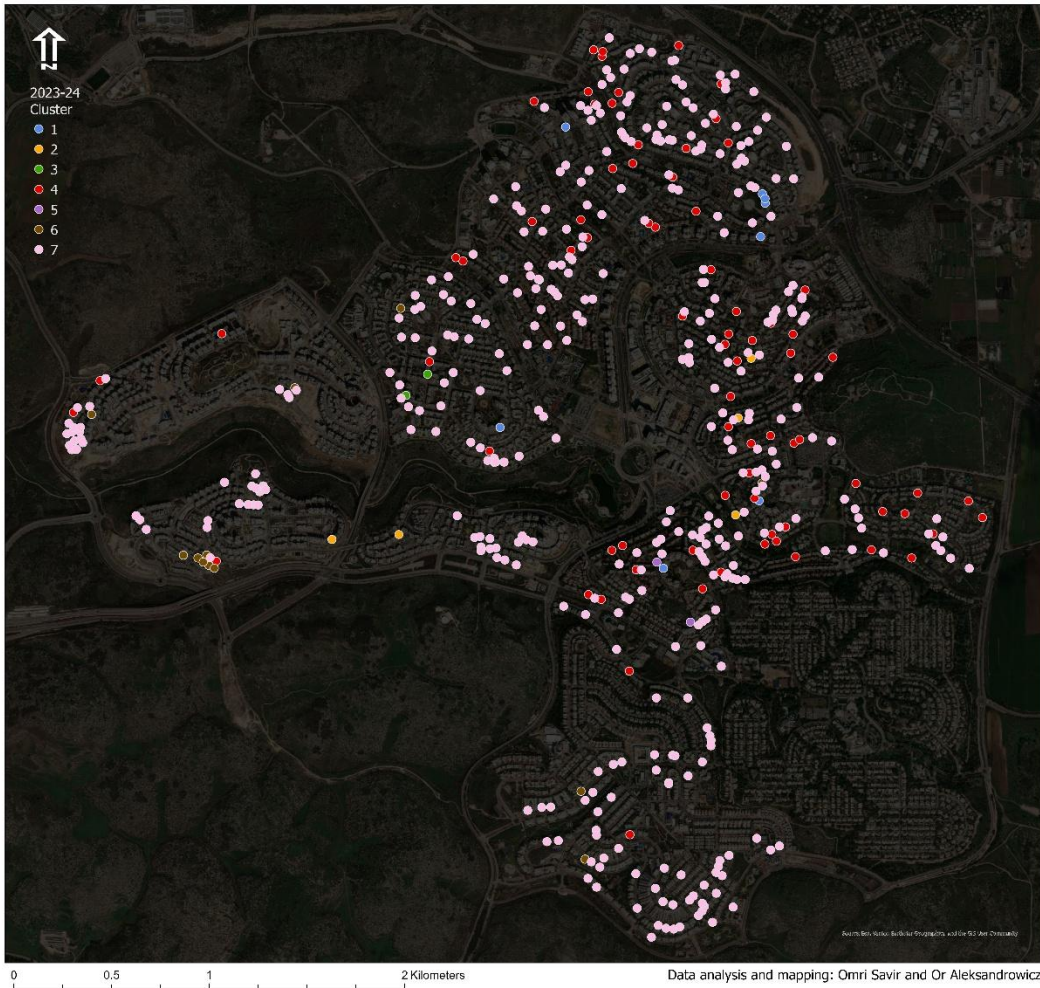


Figure 34: Geographic distribution of the K-Means clustering of the dataset buildings in Modiin.



Figure 35: Geographic distribution of the K-Means clustering of the dataset buildings in Netanya.



Figure 36: Geographic distribution of the K-Means clustering of the dataset buildings in Rishon LeZion.

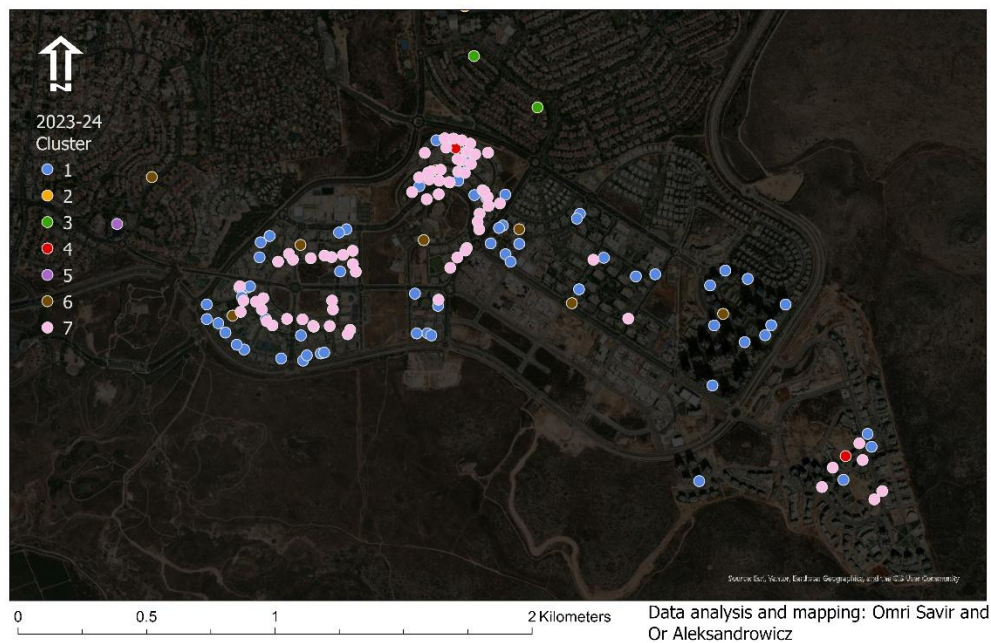


Figure 37: Geographic distribution of the K-Means clustering of the dataset buildings in Rosh Ha'ayin.

Table 10 reports the distribution of buildings across the 7 clusters. Three large clusters, 1, 4, and 7, together account for approximately 75% of all buildings, while the remaining 4 clusters (2, 3, 5, and 6) are considerably smaller.

Table 10: Distribution of buildings across the 7 K-Means clusters

Cluster	Buildings	Percent
1	1,411	23.92%
2	561	9.51%
3	416	7.05%
4	1,799	30.50%
5	111	1.88%
6	439	7.44%
7	1,162	19.70%

Several patterns emerge when comparing cluster composition across cities. Modiin is highly homogeneous: most of its buildings belong to Cluster 7, whereas in the other cities only a small share of buildings belongs to this cluster. This may partly reflect climatic differences between Modiin and the other cities, as well as differences in building characteristics. In Rosh Ha'ayin, most buildings belong to either Cluster 1 or Cluster 7. Most buildings in Cluster 4 come from Ashdod, Hadera, and Rishon LeZion. Cluster 2 is dominated by buildings from Netanya and Beer Sheva. As shown in Section 2.3.3, the consumption patterns of these 2 cities differ from other cities in the dataset. Clusters 5 and 6 are represented more strongly in Ashdod than in any of the other cities. Cluster 3 contains relatively few buildings overall, and within it, Rishon LeZion and Netanya are represented more strongly than other cities.

To characterise the consumption behaviour of each of these clusters, the same Prophet decomposition approach used in Section 2.5.1 was applied to the mean consumption series

of each cluster. Figure 38, Figure 39, and Figure 40 present the trend, weekly seasonality, and daily seasonality components, respectively, with one line per cluster. The decomposition shows that mean consumption in Cluster 5 is substantially higher than in all other clusters. Cluster 2 shows the lowest mean consumption and also shows comparatively small differences in consumption across days of the week and hours of the day. Cluster 3 shows a significant increase in consumption in summer 2024 compared with the same period in 2023. In Cluster 5, consumption peaks at noon, while in the other clusters, it peaks in the late afternoon. In clusters 5 and 2, consumption on weekends is lower than on weekdays, while in the other clusters, consumption on weekends is higher.

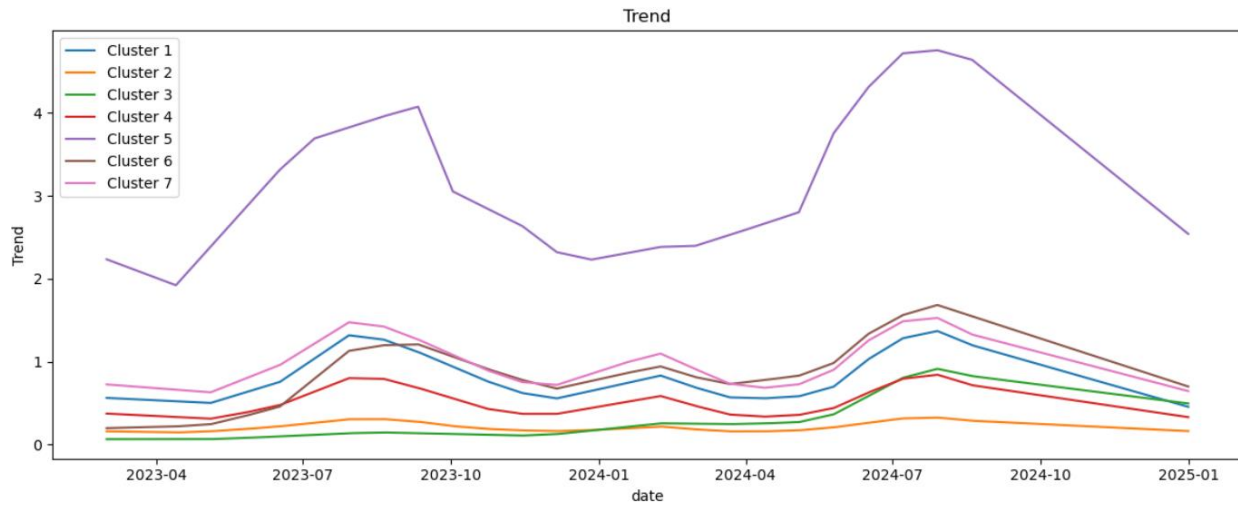


Figure 38: Prophet decomposition by cluster – trend component.

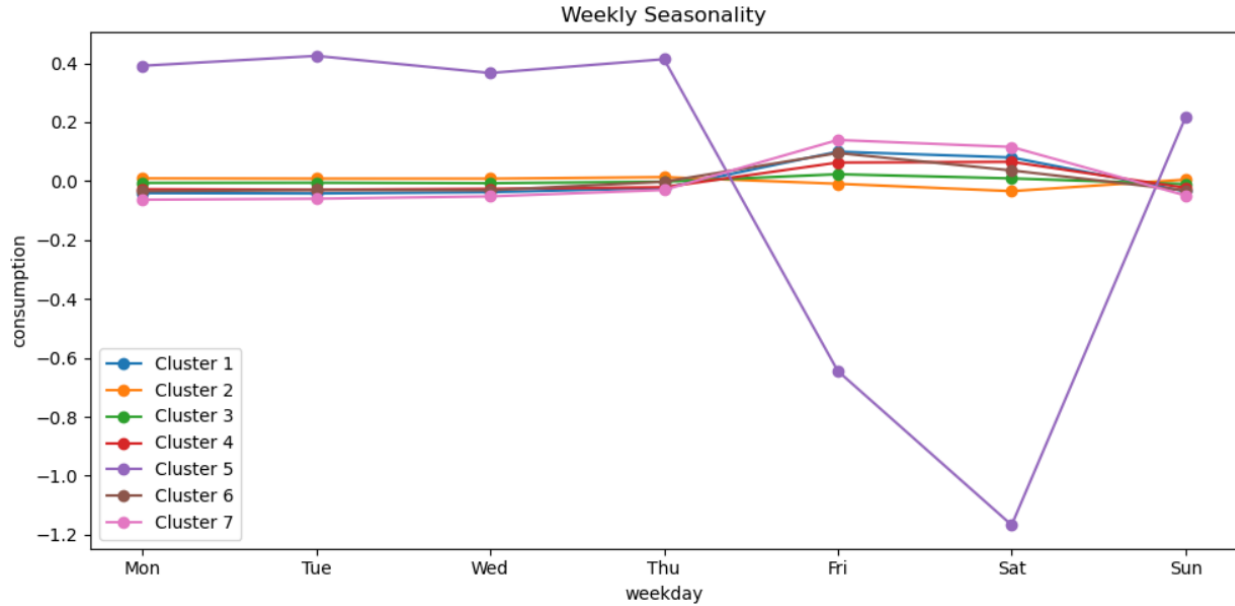


Figure 39: Prophet decomposition by cluster – weekly seasonality component.

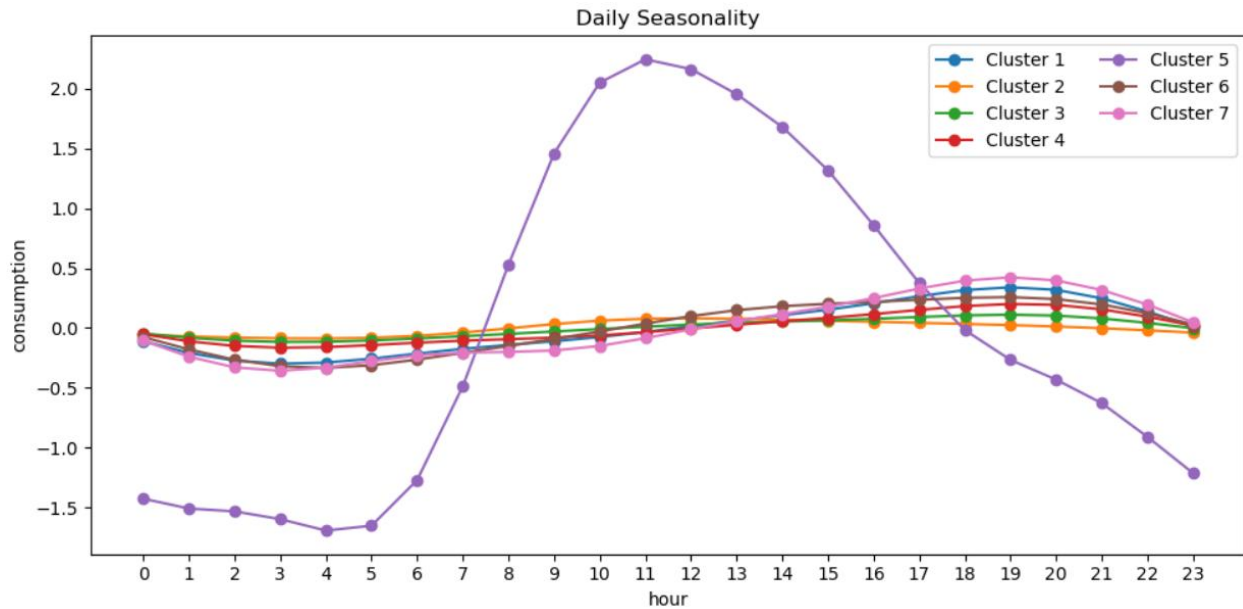


Figure 40: Prophet decomposition by cluster – daily seasonality component.

Figure 41 presents the distribution of mean normalised consumption by cluster. The results are consistent with the Prophet decomposition: Cluster 5 has the highest consumption, and Cluster 2 the lowest. Clusters 3 and 4 also show relatively low consumption compared with

Clusters 1, 6, and 7. Cluster 6 shows considerably greater variance than the other clusters, except for Cluster 5.

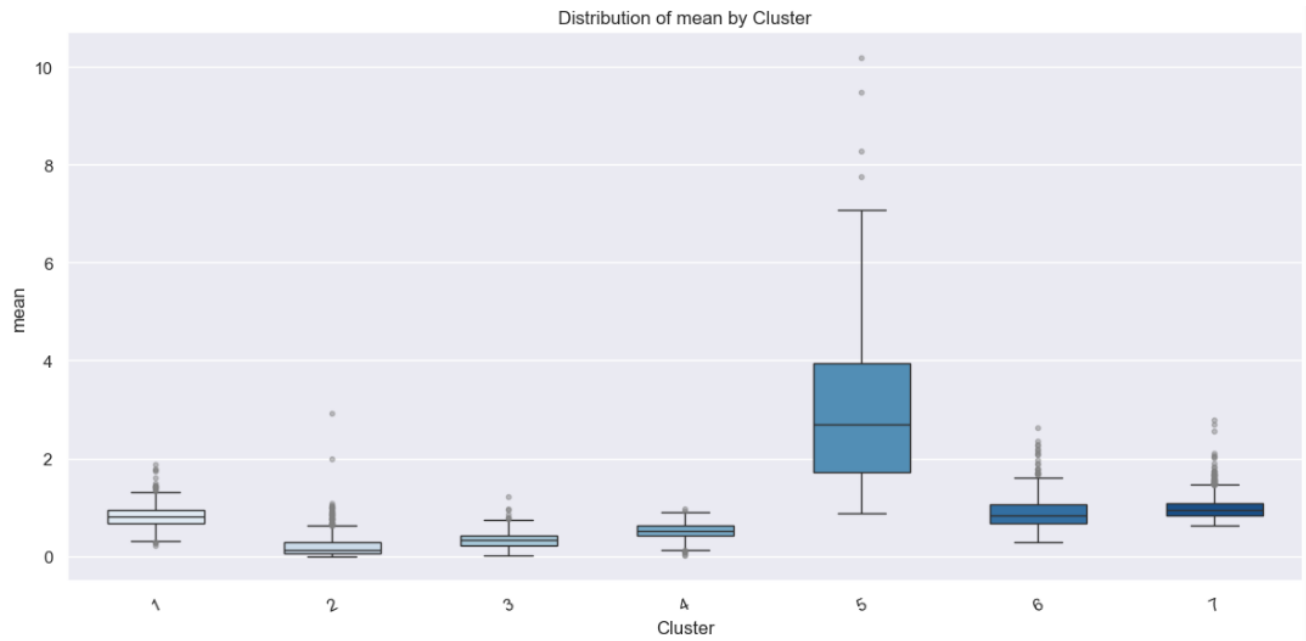


Figure 41: Distribution of mean normalised consumption by cluster (box plot). Consumption figures are average hourly kWh per meter (apartment).

Figure 42 and Figure 43 show the distributions of the Friday ratio and Saturday ratio by cluster. While most clusters show higher weekend consumption than weekday consumption, Cluster 5 shows substantially lower weekend consumption. Cluster 2 also contains a non-negligible number of buildings with lower weekend-than-weekday consumption. This aligns with the patterns in Figure 44 and Section 2.3.2, where Ashdod, Beer Sheva, and Netanya showed lower weekend-than-weekday consumption on average, and Clusters 5 and 2 are more common in these cities than in the others.

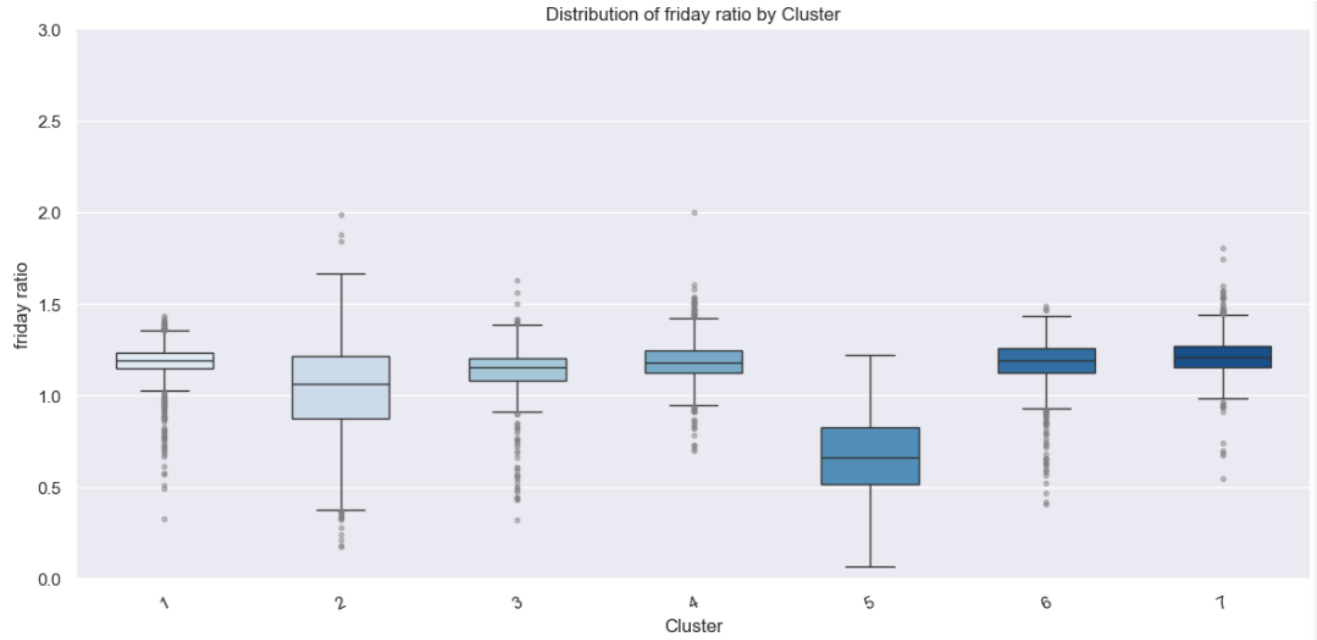


Figure 42: Distribution of the Friday ratio by cluster (box plot).

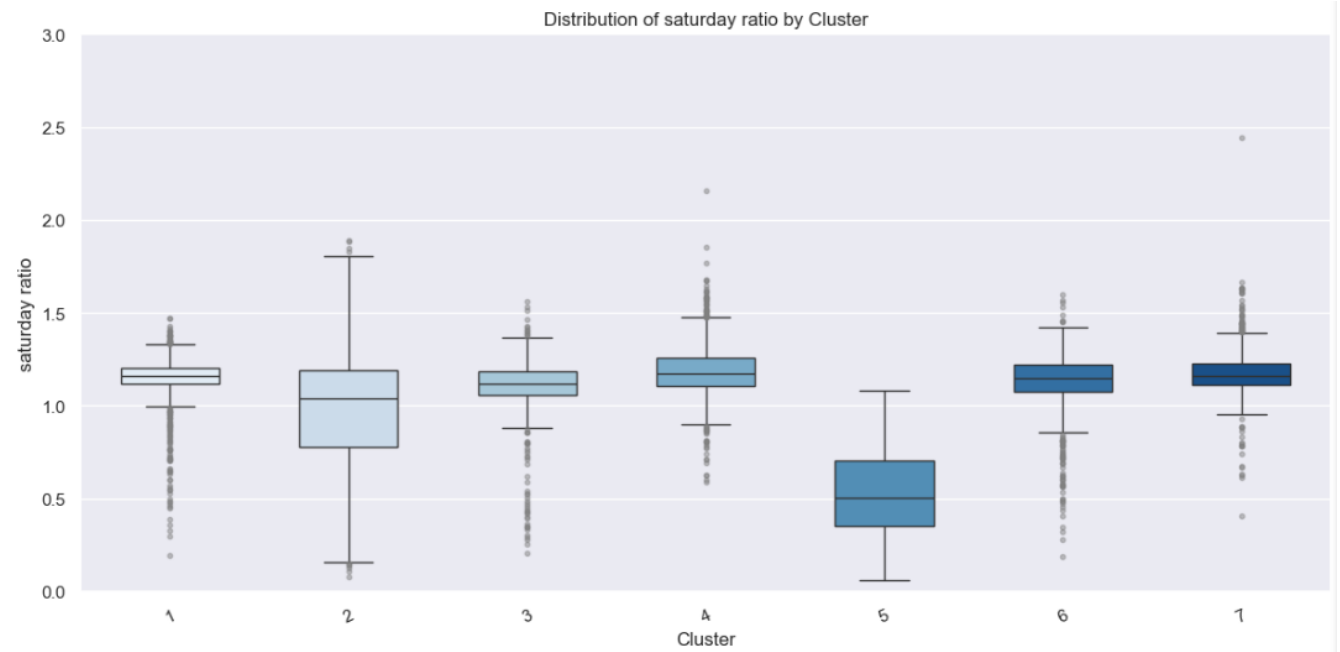


Figure 43: Distribution of the Saturday ratio by cluster (box plot).

Figure 44 and Figure 45 show the distributions of peak and minimum consumption hours by cluster. While most clusters peak in the afternoon hours, Cluster 5 peaks around midday. In Cluster 2, the median peak hour is also before 16:00, and many of its buildings peak at

midday as well. Cluster 6 also has several buildings with an earlier peak hour and shows greater variance in peak hour than the other clusters. Regarding the minimum consumption hour, Cluster 5 reaches its minimum in the early morning hours, later than the other clusters. In Cluster 1, the minimum consumption hour is earlier than in most other clusters (around 04:00), with relatively low variance. Clusters 2 and 3 also include many buildings whose minimum consumption hour falls in the early morning.

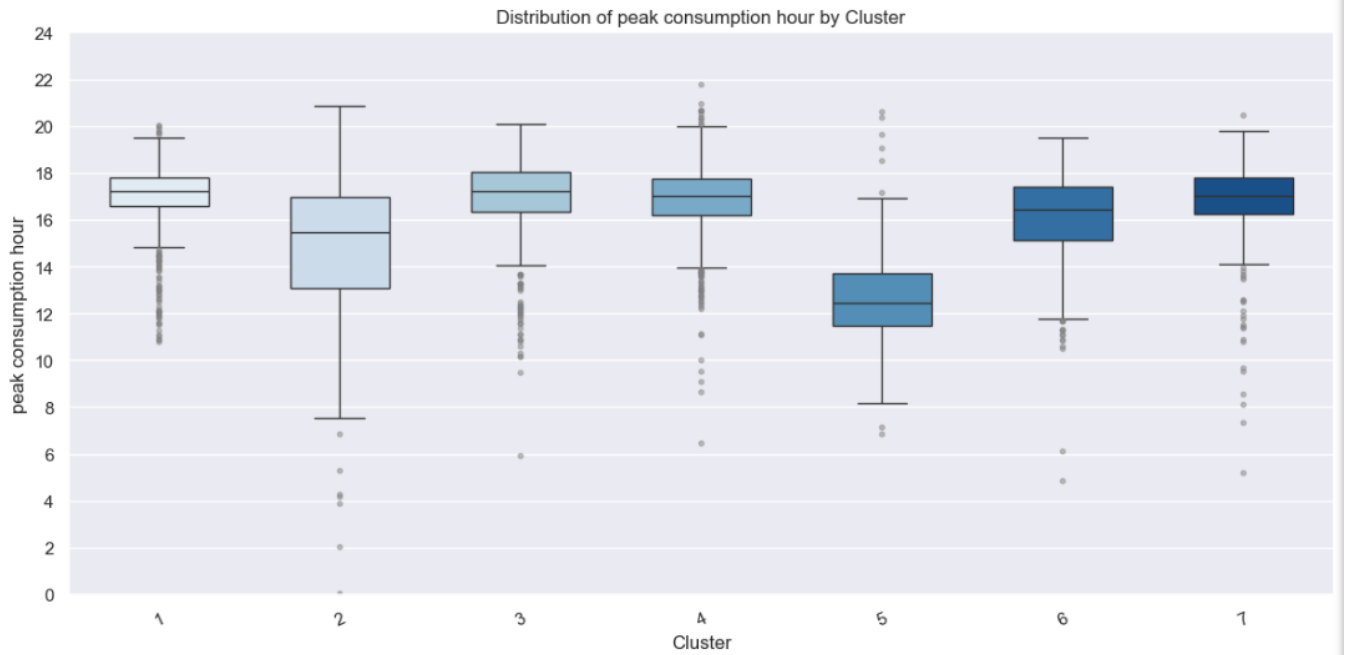


Figure 44: Distribution of the peak consumption hour by cluster (box plot).

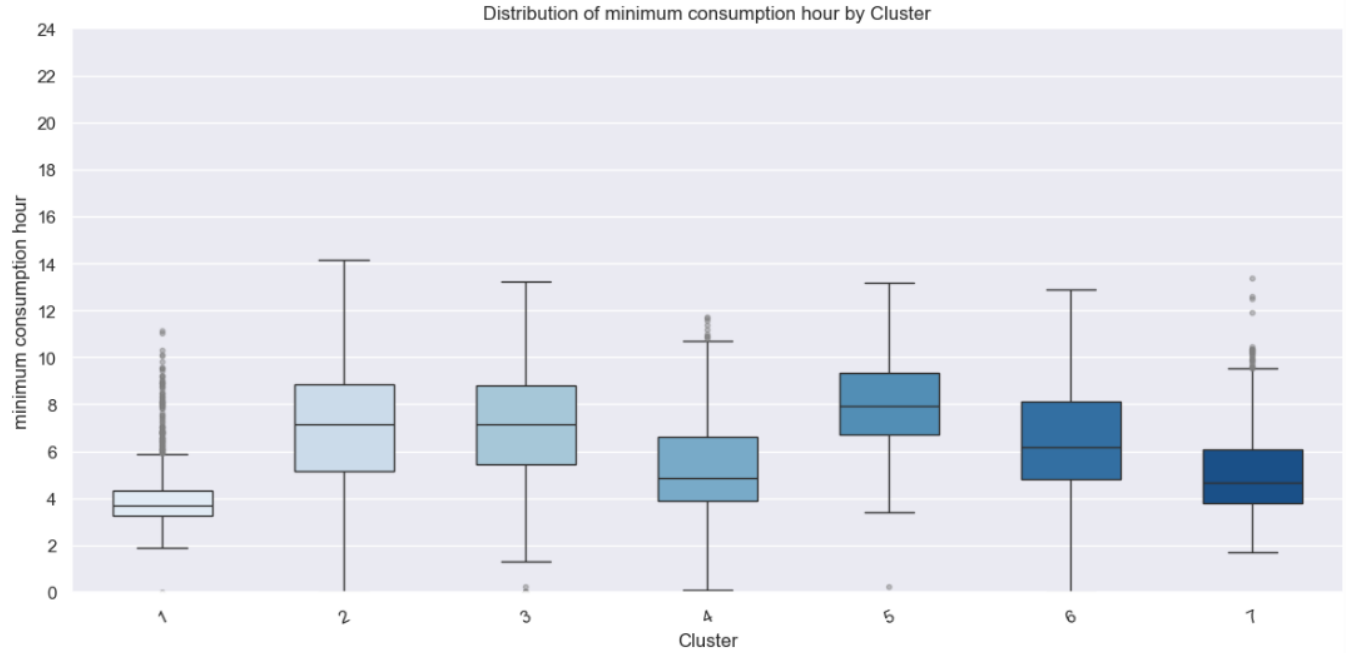


Figure 45: Distribution of the minimum consumption hour by cluster (box plot).

Figure 46 shows the distribution of the summer ratio by cluster. The differences between clusters are smaller here than for the other variables, with most buildings falling between 1.5 and 2 regardless of cluster. Cluster 3 shows a somewhat higher summer ratio than the other clusters, while Cluster 5 shows a somewhat lower ratio.

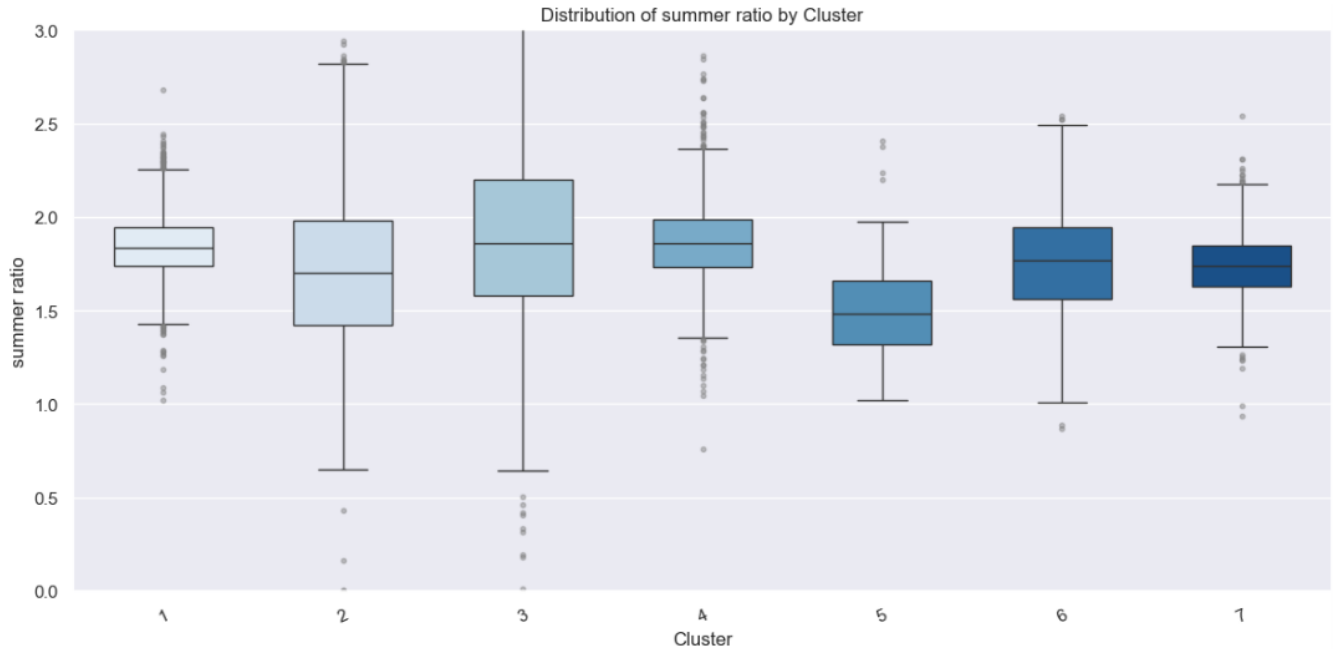


Figure 46: Distribution of the summer ratio by cluster (box plot).

Finally, Figure 47 presents a Likert-style chart comparing the distribution of socioeconomic clusters across the 7 clusters. The socioeconomic cluster is a property of the statistical zone where a building is located, not the building itself. In Cluster 7, most buildings belong to socioeconomic clusters 8–9, consistent with the fact that most buildings in this cluster are located in Modiin, where all neighbourhoods belong to socioeconomic clusters 8 or 9. Cluster 6 shows a high share of socioeconomic cluster 2 and other relatively low socioeconomic clusters; in this study's data, socioeconomic cluster 2 neighbourhoods are mostly ultra-Orthodox neighbourhoods, primarily in Ashdod. Clusters 1 and 3 show a relatively high representation of socioeconomic cluster 7, while Cluster 4 shows a relatively high representation of socioeconomic cluster 6. Clusters 5 and 2 show a comparatively diverse distribution of socioeconomic clusters, ranging from 2 to 8. The socioeconomic cluster distribution in the 3 largest clusters (1, 4, and 7) may help explain the findings in Figure 38 and Figure 41, where consumption in Cluster 7 is highest, consumption in Cluster 1 is intermediate, and consumption in Cluster 4 is lowest. This ordering appears to mirror the relative socioeconomic ranking of the clusters, suggesting a possible positive relationship between socioeconomic cluster and electricity consumption. The opposite

pattern emerges in Cluster 6, where most buildings belong to low socioeconomic clusters, yet consumption remains high (comparable in magnitude to Clusters 1 and 7, where most buildings belong to socioeconomic cluster 7 or above). This likely reflects the high concentration of the ultra-Orthodox population in this cluster, where household size is typically much larger than in the general population, driving higher electricity demand despite lower socioeconomic status.

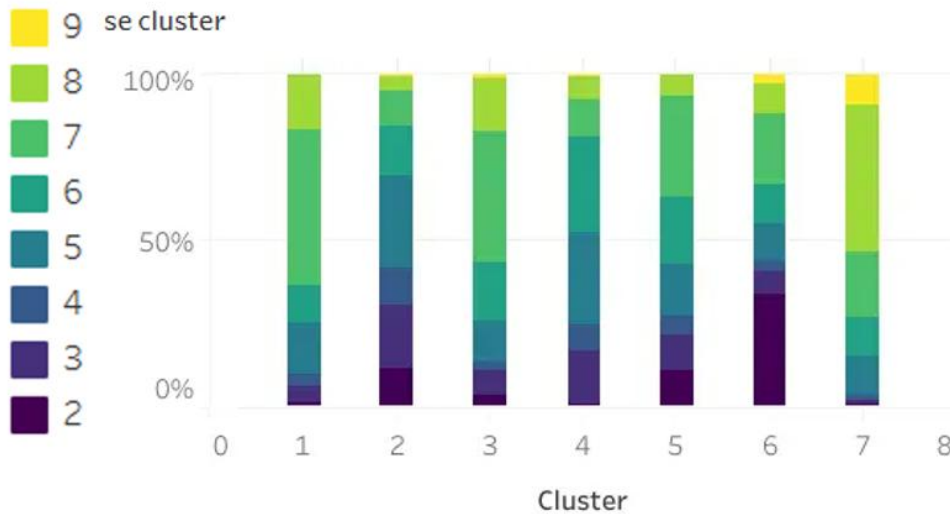


Figure 47: Distribution of socioeconomic cluster by K-Means cluster (Likert-style diverging stacked bar chart).

To assess whether the construction-consumption nexus types identified by clustering are stable over time, we repeated the clustering process described at the beginning of this section independently on 2 additional time windows: March 2023 to December 2023, and January 2024 to December 2024. The 2023-only run yielded 7 clusters, while the 2024-only run yielded 8 clusters, after the same small-cluster merging procedure.

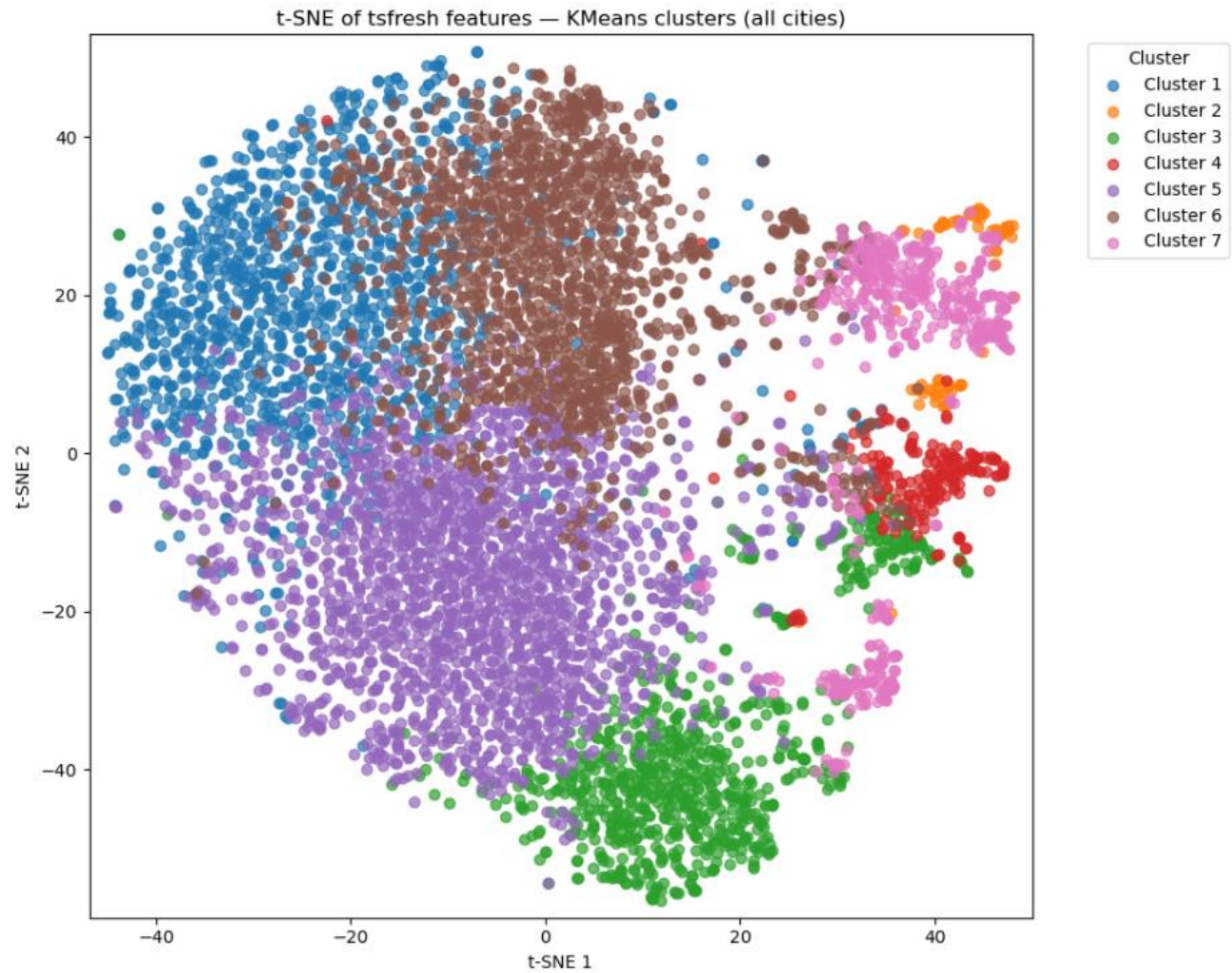


Figure 48: t-SNE projection of the tsfresh feature space, coloured by K-Means cluster assignment, 2023 only.

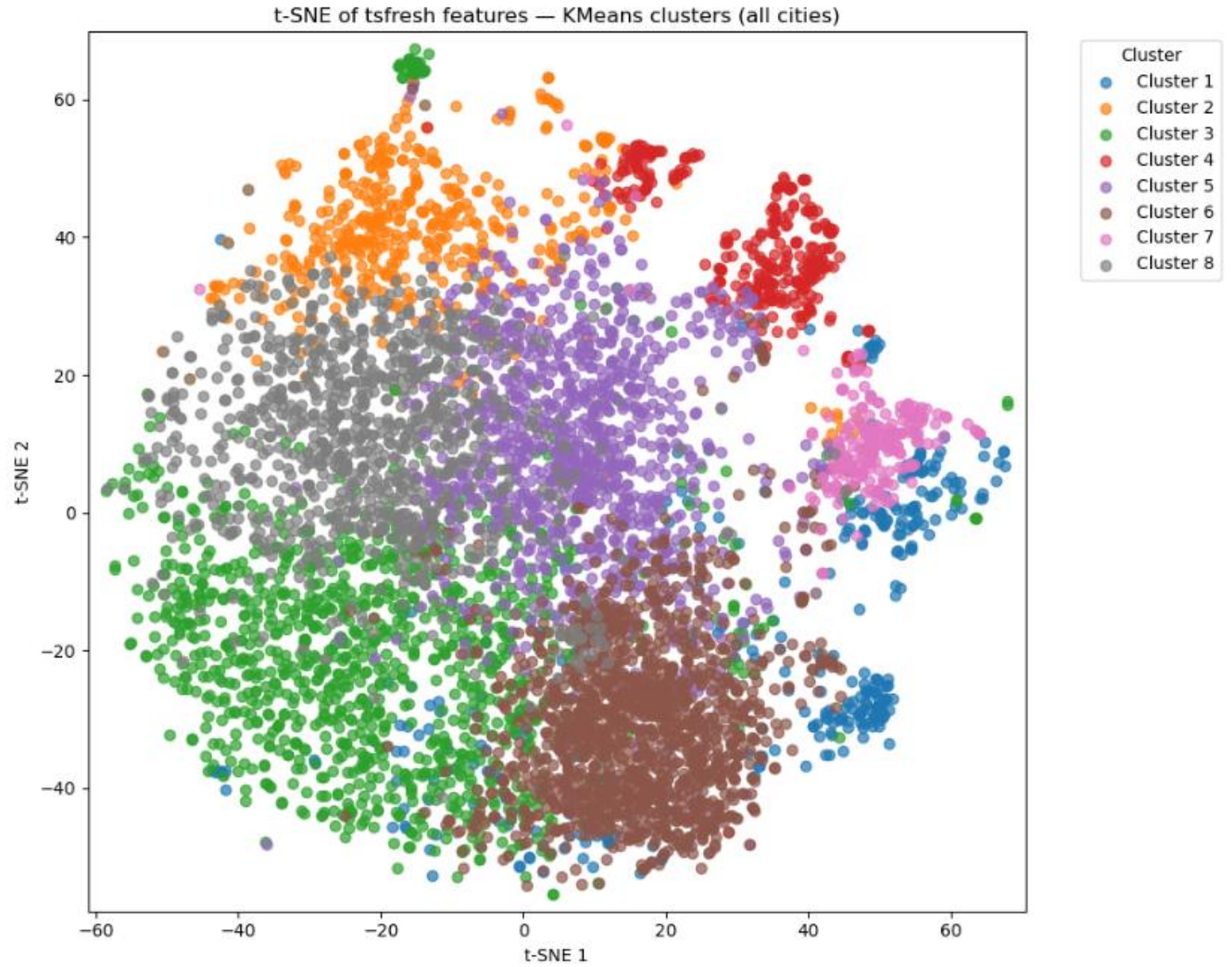


Figure 49: t-SNE projection of the tsfresh feature space, coloured by K-Means cluster assignment, 2024 only.

To examine how buildings moved between clusters across years, we constructed a cross-tabulation where rows represent the cluster a building was assigned to in 2023, and columns represent the cluster it was assigned to in 2024, analogous to a confusion matrix. Figure 50 presents this cross-tabulation in 2 forms: in absolute counts, and normalised by row (i.e., as the proportion of each 2023 cluster's buildings that were assigned to each 2024 cluster).

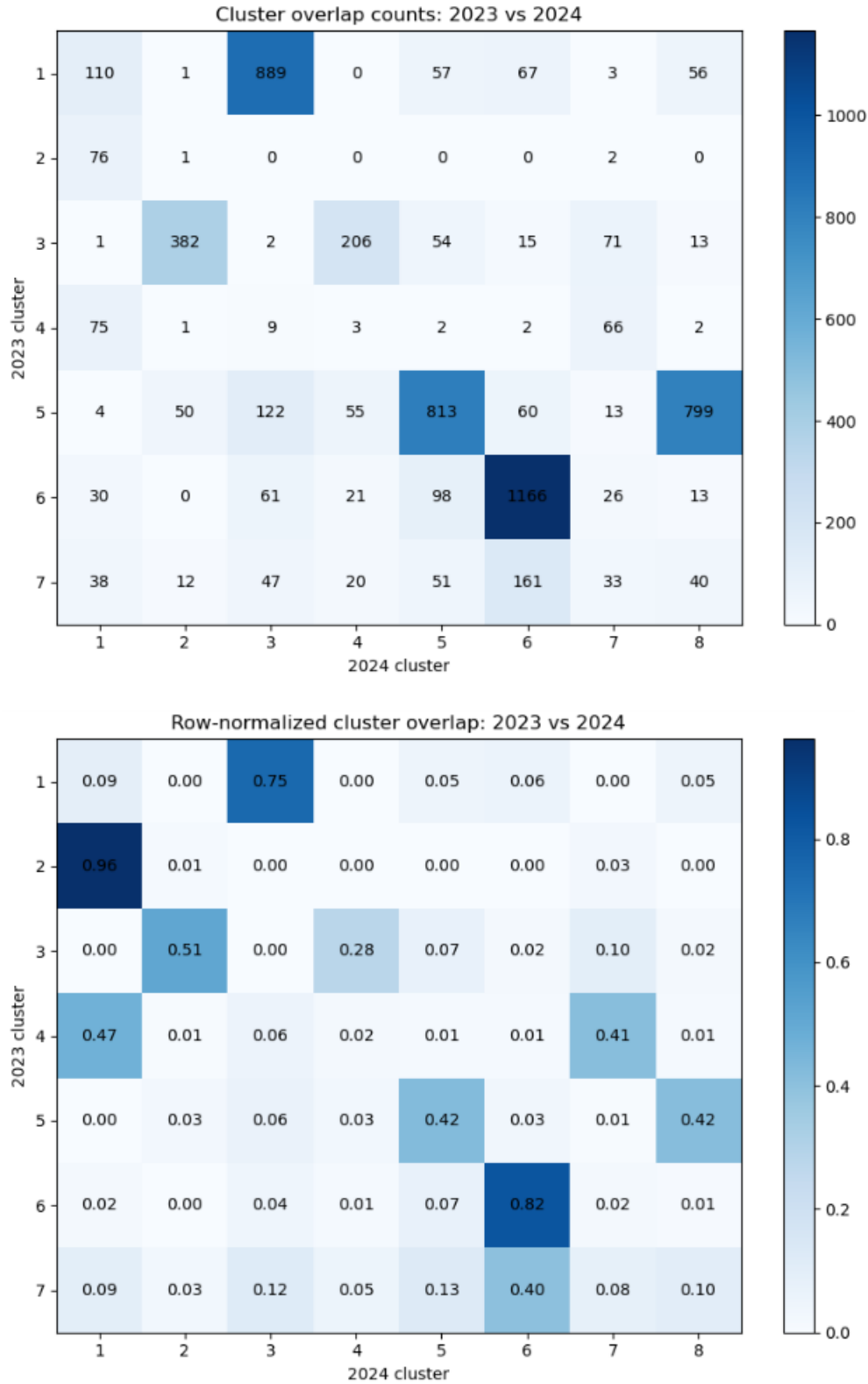


Figure 50: 2023-to-2024 cluster transition matrix. Top: absolute building counts. Bottom: row-normalised (proportion of each 2023 cluster's buildings by 2024 cluster assignment).

Some 2023 clusters are highly persistent: most buildings in 2023 Clusters 1, 2, and 6 were assigned to the same cluster in 2024. This is most striking for 2023 Cluster 2, of which 96% of buildings were assigned to 2024 Cluster 1. In contrast, several 2023 clusters effectively split across two 2024 clusters, including Clusters 3, 4, and 5. For 2023 Cluster 7, 40% of buildings were assigned to 2024 Cluster 6, while the remainder were distributed across the other 2024 clusters. Taken together, these results suggest that while the data identify certain consumer types, cluster membership is not highly stable over time for all consumer types.

To quantify this continuity more precisely, we computed a pairwise same-cluster retention score. For every pair of buildings that belonged to the same cluster in 2023, this score measures the fraction of such pairs that also belonged to the same cluster in 2024. Let n_{ij} denote the number of buildings assigned to cluster i in 2023 and cluster j in 2024, and let $N_i = \sum_j n_{ij}$ denote the total number of buildings assigned to cluster i in 2023. For a count of buildings x , the number of distinct pairs that can be formed is given by the binomial coefficient:

$$C_2(x) = \frac{x(x-1)}{2}$$

The overall pairwise retention score R is then defined as the number of building pairs that remained in the same cluster in both years, divided by the number of building pairs that were in the same cluster in 2023:

$$R = \frac{\sum_{i,j} C_2(n_{ij})}{\sum_i C_2(N_i)}$$

An analogous score R_i can be computed separately for each 2023 cluster i , restricting both the numerator and denominator to pairs originating from that cluster:

$$R_i = \frac{\sum_j C_2(n_{ij})}{C_2(N_i)}$$

Note that a random division into 8 clusters in 2024 would have resulted in an expectation of $R=0.125$.

The overall pairwise retention score was $R = 48.04\%$, indicating that roughly half of all building pairs that shared a cluster in 2023 still shared a cluster in 2024. Table 11 reports the retention score separately for each 2023 cluster.

Table 11: Pairwise same-cluster retention score by 2023 cluster

2023 Cluster	Pairwise Retention Score
1	58.08%
2	92.53%
3	35.45%
4	39.01%
5	36.02%
6	68.65%
7	21.72%

Consistent with the transition matrix, Cluster 2 has the highest retention score by far (92.53%), and Cluster 6 also shows relatively high retention (68.65%). Clusters 3, 4, and 5 have the lowest retention scores, consistent with their tendency to split across multiple 2024 clusters, while Cluster 7 has the lowest retention score overall (21.72%).

2.4 Cross-building prediction models

To develop an electricity consumption prediction model, we first tested whether a building's electricity consumption can be predicted from the consumption patterns of other buildings in a similar geographic location and under similar weather conditions. Unlike a purely temporal approach, the target building provides no training data; the model learns from other buildings and predicts consumption for buildings it did not see during training. All models in this section were trained and evaluated on the summer period only (June–

September) of each year. We analysed a subset of buildings with data in both 2023 and 2024, plus Modiin in 2022, enabling direct year-to-year comparison.

2.4.1 Model description

Two main prediction models were developed:

- Linear Regression (LR) – a simple, interpretable model that assumes a linear relationship between variables, enabling direct quantification of the effect of each variable on consumption.
- Explainable Boosting Machine (EBM) – a gradient boosting-based machine learning model that provides both global explanations of variable importance and local explanations for individual predictions (Lou et al., 2013).

We evaluated model performance at 2 complementary scales. At the building level, we compute R^2 separately for each building over all its hourly test observations, then average across buildings. At the city level, we average the hourly predictions for all buildings in a city to form a single citywide mean prediction for each hour, then compute R^2 on this aggregate series. City-level aggregation typically yields higher accuracy than the building level, as the mean cancels out idiosyncratic building-level errors.

2.4.2 Training scheme – 5-fold cross-validation

Evaluation was performed using 5-fold cross-validation. We randomly partitioned the buildings in each city into 5 equal-sized folds of approximately 20% each. In each of the 5 iterations, 4 folds (80% of buildings) serve as the training set and the remaining fold (20%) serves as the test set, so that every building receives exactly one out-of-sample prediction over the full procedure. The fold assignment is at the building level: once a building is assigned to a fold, all of its hourly observations belong to that fold. This ensures the model is never tested on hourly records from a building it saw during training. We performed the split independently for each city and year.

2.4.3 Input features

For both the LR and EBM models, 4 categories of input features were used:

- Time: day of week, hour of day, and month.
- Weather: the hour's Heat Index together with 7 daily-average lagged Heat Index features representing the mean Heat Index over each of the 7 preceding 24-hour windows.
- Consumption: 7 lagged daily-average consumption features.
- Building characteristics: statistical zone and number of floors.

For the Linear Regression model, we encoded the time axis by fitting a separate model per time slot. Each observation was assigned to one of 15 time-slot categories formed by crossing 3 day-type groups (regular weekday, Friday, Saturday) with 5 hour-of-day groups (night: 22–06 h; morning: 07–10 h; day: 11–16 h; evening: 17–21 h; late evening: 22–24 h). We fitted a separate LR model for each time slot, including month as a one-hot covariate. The EBM model received month, hour of day, and day of week directly as numeric features without sub-model splitting. In addition, the Heat Index was clipped to positive deviations above a comfort threshold of 17°C (i.e. $\max(\text{HI} - 17, 0)$), so that only thermally-driven summer excess demand is captured.

2.5 Forecasting: Prophet Model

The second prediction model we developed attempted to forecast future consumption based on past consumption data. Because we had limited access to multi-annual datasets, we tested our ability to forecast electricity consumption 2 months ahead without using recent consumption data as input. The model was designed for autonomous summer-season forecasting, primarily to support infrastructure planning and future load management. As with the cross-building prediction models, we restricted the analyses to buildings with data for both 2023 and 2024.

2.5.1 Model description

To forecast electricity consumption for each building i at an hourly resolution, we employed the Prophet model (Taylor & Letham, 2018), an additive time series forecasting framework introduced by Facebook, designed to be both scalable and interpretable. We treated each

building as a separate time series and trained an independent model per building. This design enables explainability at the individual-building level, allowing us to interpret how model components and regressors affect each building's consumption. We evaluated model performance at 2 complementary scales: the building level and the city level. For comparison, we also generated a single-series model based on the city-wide average consumption. We selected June–July 2023/2024 as the training period, August 1–10, 2023/2024 as the validation period, and August 11–September 30, 2023/2024 as the test period.

The model decomposes the observed time series into trend, seasonal, and regressor components, defined as:

$$y_i(t) = g_i(t) + s_i(t) + \sum_{j=0}^{d_i} \beta_{ij} \tilde{x}_j(t) + \varepsilon_i(t)$$

where:

- $y_i(t)$: observed electricity consumption at time t for building i (hourly resolution)
- $g_i(t)$: trend component
- $s_i(t)$: seasonal component (daily and weekly)
- $\tilde{x}_j(t)$: standardised external regressors
- β_{ij} : regression coefficients
- $d_i \in \{0, \dots, 7\}$: number of lag days included for building i
- $\varepsilon_i(t)$: stochastic error term

2.5.2 Trend component

The long-term trend $g_i(t)$ is modelled as a piecewise linear function that lets the growth rate change at automatically selected changepoints. The function is defined as:

$$g_i(t) = (k_i + a_i(t)^\top \delta_i) \cdot t + (m_i + a_i(t)^\top \gamma_i)$$

where:

- k_i : the initial rate of growth (slope)
- m_i : the initial offset (intercept)
- S : the number of potential changepoints selected prior to model fitting
- $a_i(t) \in \{0,1\}^S$: a binary indicator vector with 1 for each changepoint time prior to t , and 0 otherwise
- $\delta_i \in \mathbb{R}^S$: the vector of adjustments to the slope at each of the S changepoints
- $\gamma_i \in \mathbb{R}^S$: the vector of corresponding adjustments to the offset to ensure continuity

The changepoint prior scale hyperparameter λ_i controls the strength of the regularisation applied to the magnitude of the slope changes in δ_i .

2.5.3 Seasonality component

Prophet models use Fourier basis functions for seasonal patterns. The seasonal component $s_i(t)$ is expressed as:

$$s_i(t) = \sum_{n=1}^N [a_n \cos(\frac{2\pi nt}{P}) + b_n \sin(\frac{2\pi nt}{P})]$$

where:

- P : the period of the seasonality ($P = 24$ for daily seasonality and $P = 168$ for weekly seasonality)
- N : number of Fourier terms
- a_n, b_n : coefficients for each term

In our setting, we included both daily and weekly seasonality components to capture intraday and weekly consumption patterns. We did not include yearly seasonality or holiday effects in the model.

2.5.4 Regressors: Heat Index and lag features

The primary external regressor is the hourly heat index at time t , denoted by $x_0(t) = \text{heat_index}(t)$.

We also construct lag-based features to capture delayed thermal effects. Each lagged feature $x_j(t)$, for $j = 1, \dots, d_i$, is defined as the average hourly heat index over a 24-hour window corresponding to the j -th day preceding time t . Specifically:

$$x_j(t) = \frac{1}{24} \sum_{h=1}^{24} \text{heat index}(t - 24 \cdot j + h - 1), \quad j = 1, \dots, d_i$$

For example, $x_1(t)$ corresponds to the average heat index over hours $t - 24$ to $t - 1$, while $x_2(t)$ covers hours $t - 48$ to $t - 25$, and so on. The number of lag days d_i is selected per building based on validation performance.

Before being incorporated into the model, all regressors $x_j(t)$ are standardised using their sample mean and standard deviation computed over the training period:

$$\tilde{x}_j(t) = \frac{x_j(t) - \mu_j}{\sigma_j}$$

where μ_j and σ_j denote the mean and standard deviation of regressor j in the training data.

2.5.5 Hyperparameter estimation

Each building i undergoes a procedure for selecting 2 hyperparameters:

- $d_i \in \{0, 1, \dots, 7\}$: the number of lag days included in the model
- $\lambda_i \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$: the changepoint prior scale, which controls the strength of regularisation on changes in the trend's slope

Instead of performing a full grid search over all possible pairs (d_i, λ_i) , we used a stepwise procedure: starting with $d_i = 0$, we searched over candidate values of λ_i . If adding an additional lag day improved the validation MSE for some λ_i , we retained it and proceeded; otherwise, the process stopped. We thus chose the final hyperparameter values per building as the combination of lag days and changepoint prior scale that minimised the validation MSE.

3. Results and discussion

3.1 Cross-Building Prediction Models

3.1.1 Target variable: normalisation strategy

Before applying models to all 7 cities, we investigated how the target variable affects prediction accuracy using Modiin data for 2022 and 2023. 3 consumption normalisation strategies were compared: (i) raw total building consumption without normalisation; (ii) consumption normalised by number of smart meters (reflecting the number of residential units); and (iii) consumption normalised by total building floor area (m^2). For each strategy, we normalised the corresponding consumption features the same way as the target variable. Table 12 and Table 13 report R^2 at the urban and building level for LR and EBM models.

In both years, predicting raw (unnormalised) consumption yielded the highest building-level accuracy by a large margin: building-level R^2 reached 0.85 (LR) and 0.84 (EBM) in 2022, compared with 0.47–0.60 for meter-normalised and 0.60–0.62 for area-normalised targets. Urban-level differences were small across all 3 strategies. Floor-area normalisation ranked second and meter-based normalisation last. Based on these results, all subsequent Modiin models use raw consumption as the target variable. For the all-cities analyses in Section 2.3.3, we use meter-normalised consumption instead.

The decision to normalise consumption by meter and not by floor area was taken because the process of calculating total floor area per building is prone to errors stemming from different sources, as explained in Section 2.1.3: we calculate building height based on a digital surface model, overlaid with a building footprints layer; we assume a standard floor height for all buildings and calculate the number of floors by dividing the gross building height by it; and we cannot separate between apartment area (which is climatized) and the non-conditioned shared public areas of each building. Each step can accumulate error relative to the actual conditioned spaces in each building. Conversely, the input IEC dataset includes the number of meters per building and can therefore be regarded as fully reliable.

Table 12: R^2 by target variable normalisation strategy, Modiin 2022 (summer, June–September)

Target variable	R^2 Urban (LR)	R^2 Urban (EBM)	R^2 Building (LR)	R^2 Building (EBM)
Without normalisation	0.891	0.974	0.850	0.841
Normalisation by building floor area	0.890	0.977	0.595	0.623
Normalisation by number of smart meters	0.895	0.986	0.470	0.511

Table 13: R^2 by target variable normalisation strategy, Modiin 2023 (summer, June–September)

Target variable	R^2 Urban (LR)	R^2 Urban (EBM)	R^2 Building (LR)	R^2 Building (EBM)
Without normalisation	0.917	0.976	0.840	0.839
Normalisation by building floor area	0.916	0.978	0.630	0.649
Normalisation by number of smart meters	0.916	0.979	0.611	0.617

3.1.2 Feature selection

Using the best-performing target (raw consumption), we evaluated each feature category's contribution by progressively removing feature groups. Starting with the full model (prediction hour + weather conditions + consumption + building characteristics), we first removed building characteristics, then removed consumption history, leaving only time and weather features. Table 14 and Table 15 report results for Modiin 2022 and 2023, respectively. While a building's age and construction technology may affect the efficiency of its energy consumption for climatisation, such data are not available at scale and accuracy for the buildings in our dataset. Therefore, this framework includes only the number of floors and a building's statistical zone, which reflects geographic proximity and shared socio-economic attributes among adjacent buildings.

The results (Table 14 and Table 15) show that the building characteristics used (number of floors, statistical zone) contribute negligibly to prediction accuracy: removing them reduces R^2 by at most 0.001 at both the building and urban level in both years. This suggests that each building's historical consumption pattern already partially embeds physical building properties, and that detailed building data are not required for accurate prediction. In light of this finding, we excluded building characteristics from all subsequent models. Removing consumption history has a more pronounced effect, particularly at the urban level: urban R^2 drops from ~ 0.97 – 0.99 to ~ 0.86 – 0.88 (EBM) when only time and weather features are used, reflecting that the consumption history of neighbouring buildings provides an important signal for predicting a given building's demand.

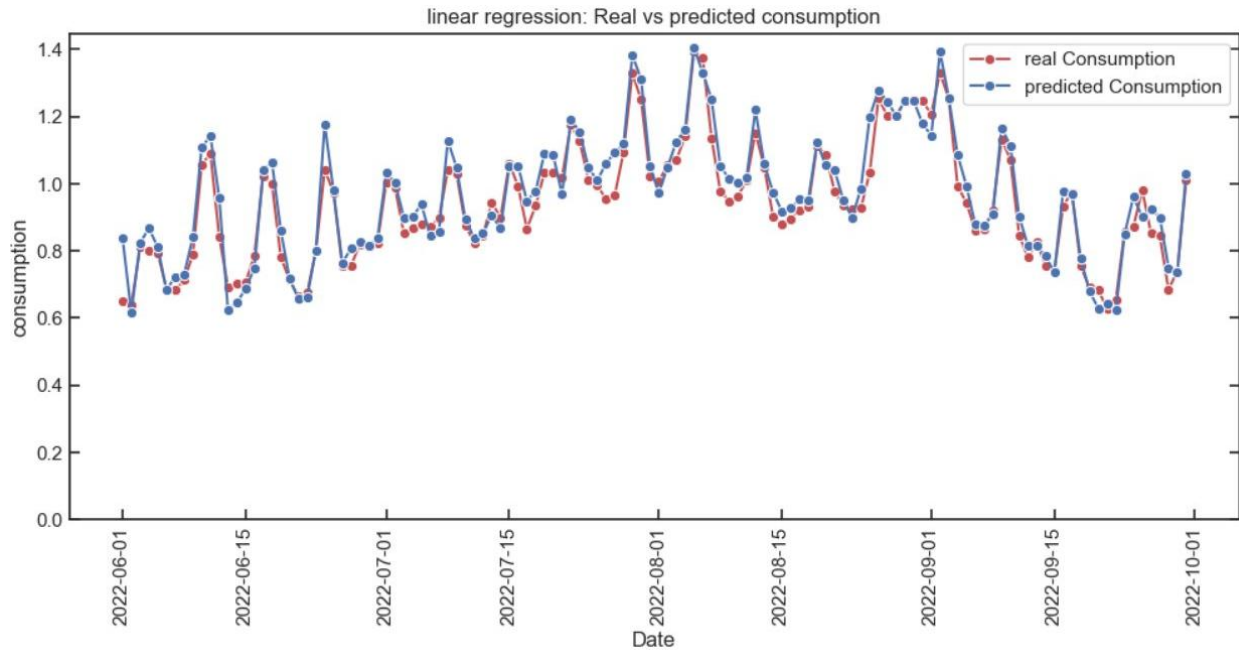
We fitted the chosen features (time + weather + consumption) for Modiin for both years, using raw consumption as the target. Figure 51 and Figure 52 plot the daily mean predicted vs observed citywide consumption for 2022 (LR and EBM, respectively), and Figure 53 and Figure 54 show the equivalent plots for 2023. In both years and for both models, the predicted series closely tracks the observed series throughout the summer.

Table 14: R^2 by input feature set, Modiin 2022 (summer, June–September, raw consumption target)

Input features	R^2 Urban (LR)	R^2 Urban (EBM)	R^2 Building (LR)	R^2 Building (EBM)
Time + Weather + Consumption + Building characteristics	0.891	0.974	0.850	0.841
Time + Weather + Consumption	0.890	0.973	0.849	0.841
Time + Weather	0.822	0.858	0.843	0.826

Table 15: R^2 by input feature set, Modiin 2023 (summer, June–September, raw consumption target)

Input features	R^2 Urban (LR)	R^2 Urban (EBM)	R^2 Building (LR)	R^2 Building (EBM)
Time + Weather + Consumption + Building characteristics	0.917	0.976	0.840	0.839
Time + Weather + Consumption	0.914	0.974	0.840	0.838
Time + Weather	0.836	0.880	0.829	0.818

**Figure 51: LR model – daily mean predicted vs. observed consumption, Modiin 2022 (summer). Consumption figures are average hourly kWh per meter (apartment).**

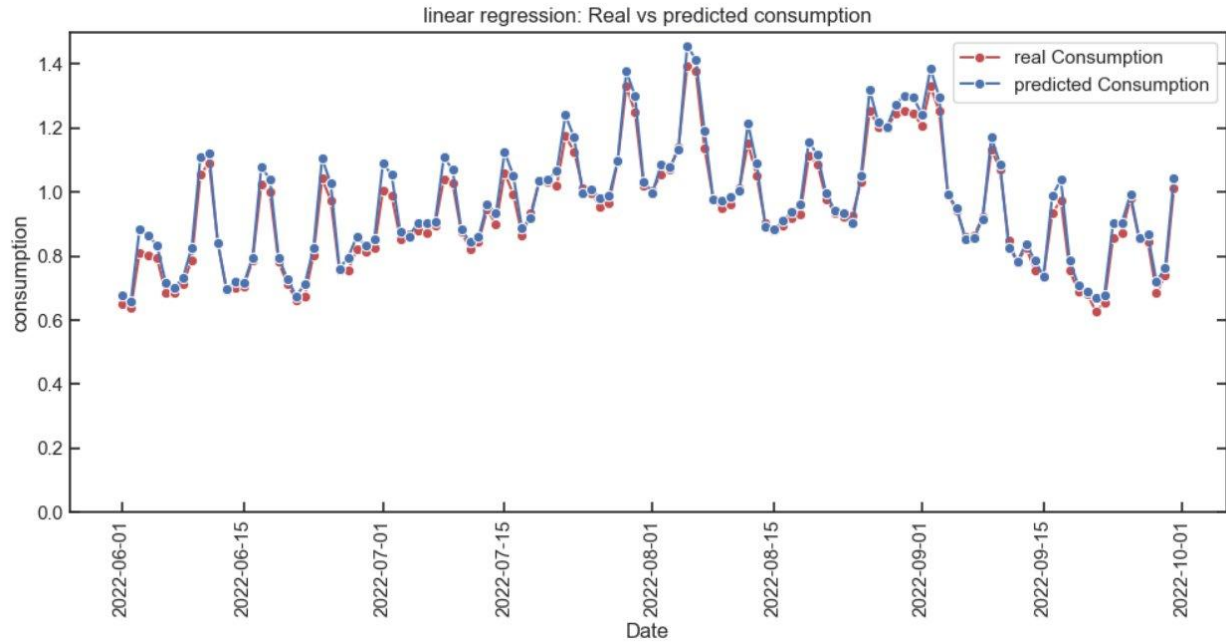


Figure 52: EBM model – daily mean predicted vs. observed consumption, Modiin 2022 (summer). Consumption figures are average hourly kWh per meter (apartment).

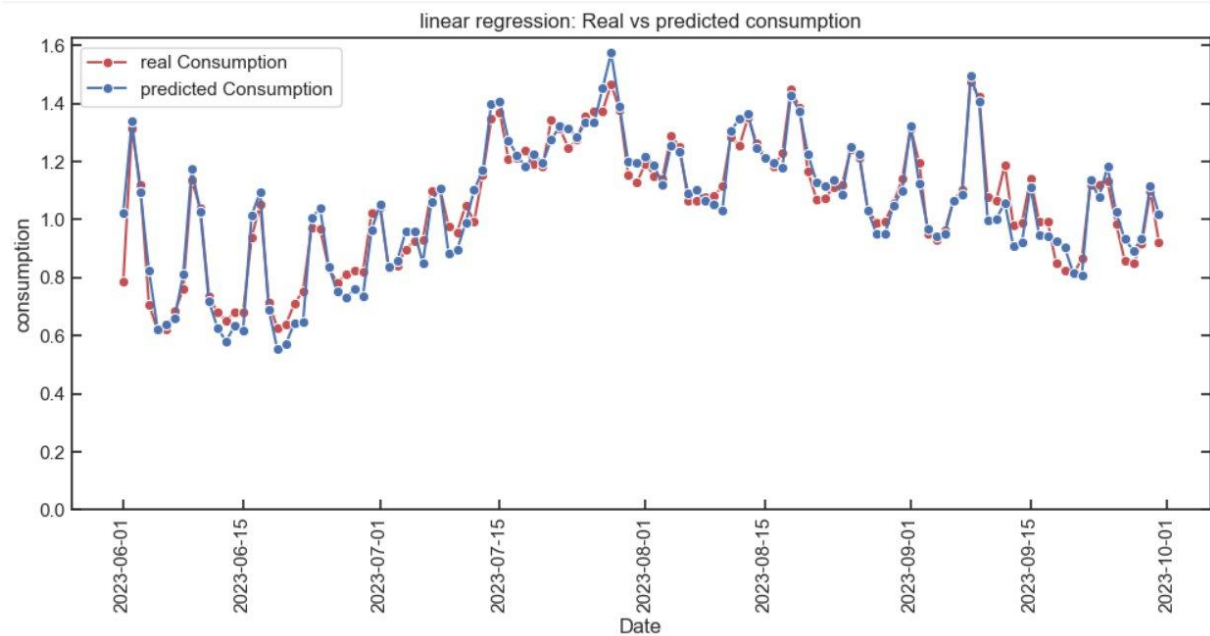


Figure 53: LR model – daily mean predicted vs. observed consumption, Modiin 2023 (summer). Consumption figures are average hourly kWh per meter (apartment).

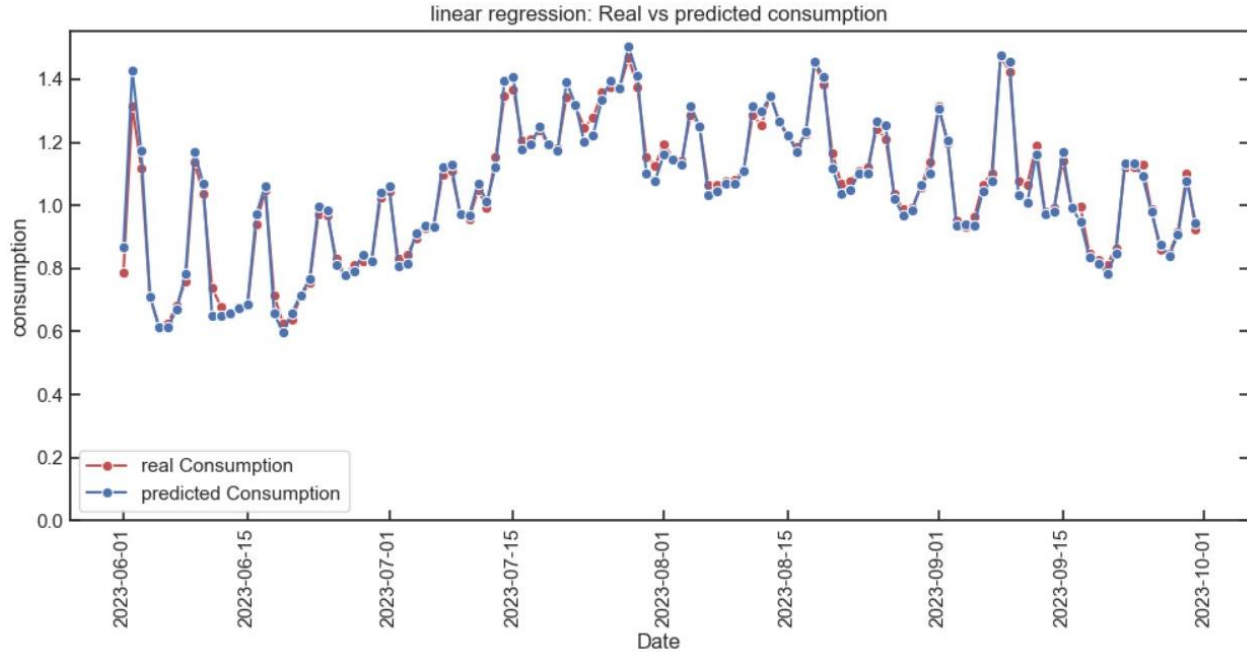


Figure 54: EBM model – daily mean predicted vs. observed consumption, Modiin 2023 (summer). Consumption figures are average hourly kWh per meter (apartment).

3.1.3 Results across all cities

Table 16 presents results across all 7 cities for 2023 and 2024. These models use time + weather + consumption features (without building characteristics) and meter-normalised consumption as the target variable, enabling consistent comparisons across cities. The models demonstrated high accuracy across all cities and both years. City-level EBM R^2 ranged from 0.981 to 0.994 in 2023 and 0.986 to 0.995 in 2024, while city-level LR R^2 ranged from 0.881 to 0.923 in 2023 and 0.819 to 0.902 in 2024. At the individual building level, accuracy was lower but still substantial: EBM building-level R^2 ranged from 0.548 to 0.713 in 2023 and 0.458 to 0.742 in 2024; LR building-level R^2 ranged from 0.494 to 0.711 in 2023 and 0.475 to 0.726 in 2024. In most cities, the LR model achieved a higher city-level R^2 in 2023 than in 2024, whereas the EBM city-level results were broadly similar across years. Hadera showed a notable drop in building-level R^2 in 2024 for both models. Beer Sheva and Rosh Ha'ayin showed slightly improved building-level accuracy in 2024.

Table 16: R^2 values for all 7 cities – LR and EBM models, 2023 and 2024 (same-buildings subset)

City	Year	Buildings	R^2 Urban (LR)	R^2 Urban (EBM)	R^2 Building (LR)	R^2 Building (EBM)
Ashdod	2023	2,452	0.923	0.988	0.697	0.653
Ashdod	2024	2,452	0.897	0.993	0.707	0.686
Beer Sheva	2023	745	0.905	0.990	0.710	0.713
Beer Sheva	2024	745	0.819	0.994	0.726	0.742
Hadera	2023	1,529	0.910	0.987	0.711	0.700
Hadera	2024	1,529	0.867	0.991	0.475	0.458
Modiin	2023	2,272	0.905	0.994	0.610	0.638
Modiin	2024	2,272	0.902	0.995	0.593	0.626
Netanya	2023	935	0.885	0.981	0.494	0.548
Netanya	2024	935	0.868	0.988	0.515	0.563
Rishon LeZion	2023	1,815	0.893	0.981	0.599	0.580
Rishon LeZion	2024	1,815	0.877	0.986	0.582	0.528
Rosh Ha'ayin	2023	271	0.881	0.992	0.669	0.627
Rosh Ha'ayin	2024	271	0.866	0.992	0.719	0.682

3.2 Forecasting – Prophet Model

3.2.1 Building-level and city-level performance

This section summarises forecasting performance at 2 aggregation levels: building level, where we evaluate predictions for individual buildings at hourly resolution across the entire test period; and city level, where we aggregate consumption across all buildings within a city into a single citywide time series and compute performance metrics. For each city, we report R^2 and symmetric Mean Absolute Percentage Error (sMAPE). Aggregating to the city level

typically reduces noise and yields higher R^2 and lower sMAPE than the building level, as idiosyncratic building-level errors tend to cancel out.

Table 17: Summary by city – building-level and city-level metrics, plus single-series comparison (2023 and 2024, same-buildings subset)

City	Year	Buildings	R^2 Bldg	sMAPE Bldg (%)	R^2 City	sMAPE City (%)	R^2 Single	sMAPE Single (%)
Ashdod	2023	2,452	0.576	32.88	0.906	7.07	0.835	9.32
Ashdod	2024	2,452	0.689	26.76	0.890	7.50	0.542	15.79
Beer Sheva	2023	745	0.691	65.89	0.923	4.98	0.288	19.21
Beer Sheva	2024	745	0.792	57.10	0.791	9.43	0.719	11.86
Hadera	2023	1,529	0.574	32.10	0.929	4.64	0.839	7.54
Hadera	2024	1,529	0.342	30.21	0.855	7.02	0.846	7.80
Modiin	2023	2,272	0.132	47.23	0.835	9.49	0.634	15.84
Modiin	2024	2,272	0.339	39.43	0.866	7.91	0.799	11.27
Netanya	2023	935	0.607	50.31	0.898	6.62	0.482	14.60
Netanya	2024	935	0.413	44.51	0.965	3.53	0.905	6.37
Rishon LeZion	2023	1,815	0.550	40.28	0.910	5.43	0.831	7.82
Rishon LeZion	2024	1,815	0.421	30.29	0.912	5.65	0.861	7.55
Rosh Ha'ayin	2023	271	0.257	36.11	0.851	6.61	0.773	8.69
Rosh Ha'ayin	2024	271	0.135	38.28	0.571	12.93	0.848	8.67

Figure 55 and Figure 56 illustrate the model's forecasts for a representative city, Ashdod, for 2023 and 2024, respectively. In both years, the predicted citywide daily consumption closely tracks the observed series, capturing both the gradual seasonal decline through September and the week-to-week fluctuations driven by short-term heat events.

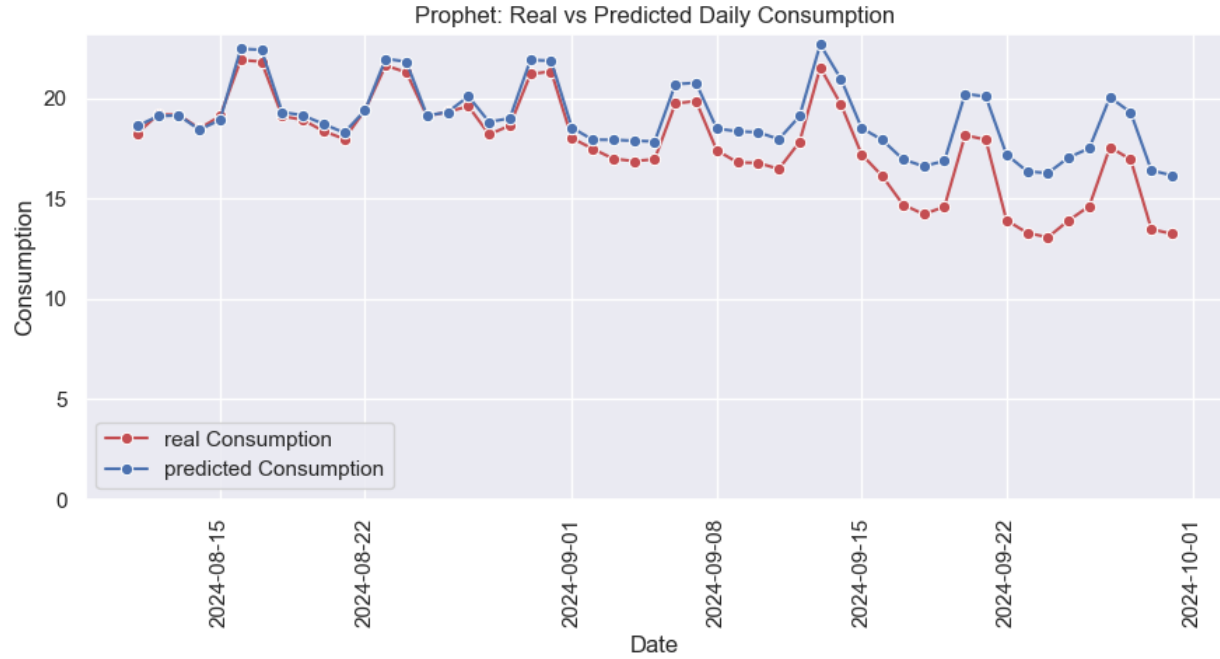


Figure 55: Prophet model – real vs. predicted daily consumption, Ashdod, 2023. Consumption figures are average daily kWh per meter (apartment).

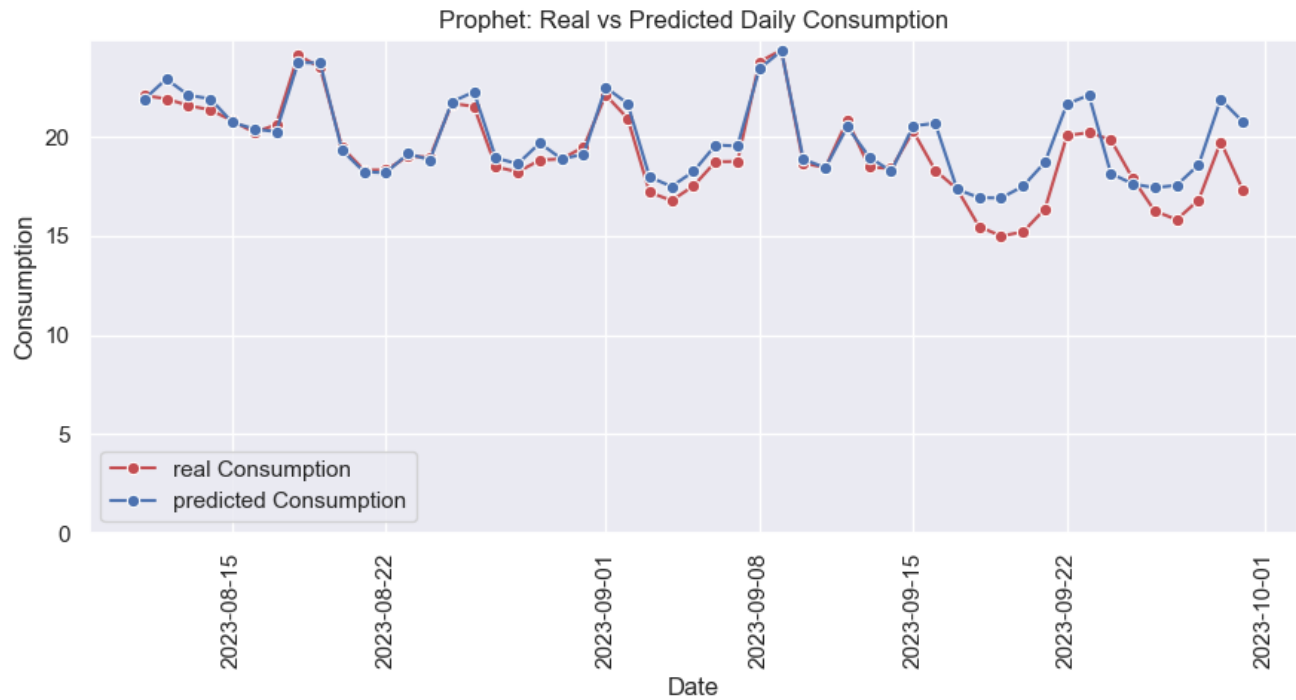


Figure 56: Prophet model – real vs. predicted daily consumption, Ashdod, 2024. Consumption figures are average daily kWh per meter (apartment).

3.2.2 Citywide single-series vs. per-building modelling

We also tested an alternative approach to fitting a separate model for each building by first averaging the hourly electricity consumption across all buildings in a city to create a single citywide series, and then training one Prophet model on this aggregate. This single-series approach follows the same structure as the per-building model, but without the building-specific index. Empirical results across all cities examined indicate that fitting separate models per building and subsequently aggregating predictions at the city level consistently outperformed the single-series citywide model (see Table 17).

3.2.3 Using forecasted weather data instead of observations

The models described above used real meteorological observations as input. However, real-world predictions, especially for managing the electricity grid, typically rely on weather forecasts issued before the prediction target time. We therefore also evaluated the model's performance using forecasted weather data instead of observed meteorological data. The Israel Meteorological Service issues one forecast run per day at 00:00, which provides hourly predictions of air temperature and relative humidity for 90 hours ahead. Consequently, each target hour has multiple available forecasts. In this study, we consider 2 forecasts for each hour: the forecast issued at 00:00 on the same calendar day (the "Same-Day forecast") and the forecast issued at 00:00 on the previous calendar day (the "Previous-Day forecast"). Using the same models, we replaced the observed weather data with these 2 forecasts as inputs. The analyses presented here use only 2023 data.

We evaluated two distinct approaches for incorporating weather forecasts instead of observed meteorological data. First, we trained models on observed weather data (as in the baseline Prophet setup) and, at prediction time, substituted the observed regressors with weather forecasts. Second, we trained models directly on the weather forecasts and also tested them using the same forecast inputs.

At the building level, replacing observed weather variables with forecasted inputs leads to a modest decline in predictive accuracy. This decline, however, is generally small and not practically significant; in fact, in 2 cities (Modi'in and Rishon LeZion), models trained directly

on weather forecasts slightly outperform those trained on observed data. Consistently, models trained on forecasts yield better results than models trained on observations when both are tested using forecasted inputs. Table 18 summarises the building-level results.

At the city level, models trained directly on weather forecasts consistently outperform those trained on observed data when evaluated with forecasted inputs, and the performance gap is substantially larger than at the building level. Except for Rosh Ha'ayin, a city with a relatively small building stock, forecast-based models performed nearly as well as those using observed weather data and, in several cities (Modiin, Netanya, and Hadera), even delivered superior accuracy. Table 19 summarises the city-level results.

Furthermore, forecasts issued at midnight on the same day are marginally more accurate than those issued at midnight the previous day, at both the building and city levels, though the difference is negligible. Overall, the results suggest that the Prophet framework remains robust under forecast substitution, with performance that stays relatively stable across cities despite the added uncertainty from forecasted weather regressors.

Figure 57 illustrates this comparison for Ashdod in 2023, showing the observed consumption alongside predictions from 3 model variants: Obs (trained and evaluated on observed weather), SD→SD (trained and evaluated on same-day forecasts), and PD→PD (trained and evaluated on previous-day forecasts). At several points, the forecast-based models (SD→SD and PD→PD) diverge from observed consumption while the Obs model remains accurate, suggesting these prediction errors stem from inaccuracies in the underlying weather forecast rather than the consumption model itself.

Table 18: Building-level R^2 with observed vs. forecasted weather, 2023 only (same-day vs. previous-day 00:00 issuances)

City	Obs	Obs→SD	Obs→PD	SD→SD	PD→PD
Ashdod	0.576	0.694	0.694	0.677	0.683
Modiin	0.132	0.401	0.406	0.448	0.447
Rishon LeZion	0.550	0.684	0.683	0.705	0.703
Beer Sheva	0.691	0.655	0.656	0.655	0.654
Hadera	0.574	0.779	0.775	0.800	0.799
Netanya	0.607	0.818	0.814	0.835	0.837
Rosh Ha'ayin	0.257	0.842	0.844	0.851	0.839

Obs = trained & tested on observed weather; Obs→SD = trained on observed, tested with Same-Day forecasts; Obs→PD = trained on observed, tested with Previous-Day forecasts; SD→SD = trained & tested on Same-Day forecasts; PD→PD = trained & tested on Previous-Day forecasts.

Table 19: City-level R^2 with observed vs. forecasted weather, 2023 only (same-day vs. previous-day 00:00 issuances)

City	Obs	Obs→SD	Obs→PD	SD→SD	PD→PD
Ashdod	0.906	0.795	0.792	0.866	0.836
Modiin	0.835	0.710	0.709	0.866	0.855
Rishon LeZion	0.910	0.663	0.654	0.892	0.881
Beer Sheva	0.923	0.614	0.618	0.882	0.870
Hadera	0.929	0.620	0.585	0.893	0.893
Netanya	0.898	0.546	0.504	0.859	0.845
Rosh Ha'ayin	0.851	0.639	0.632	0.782	0.747

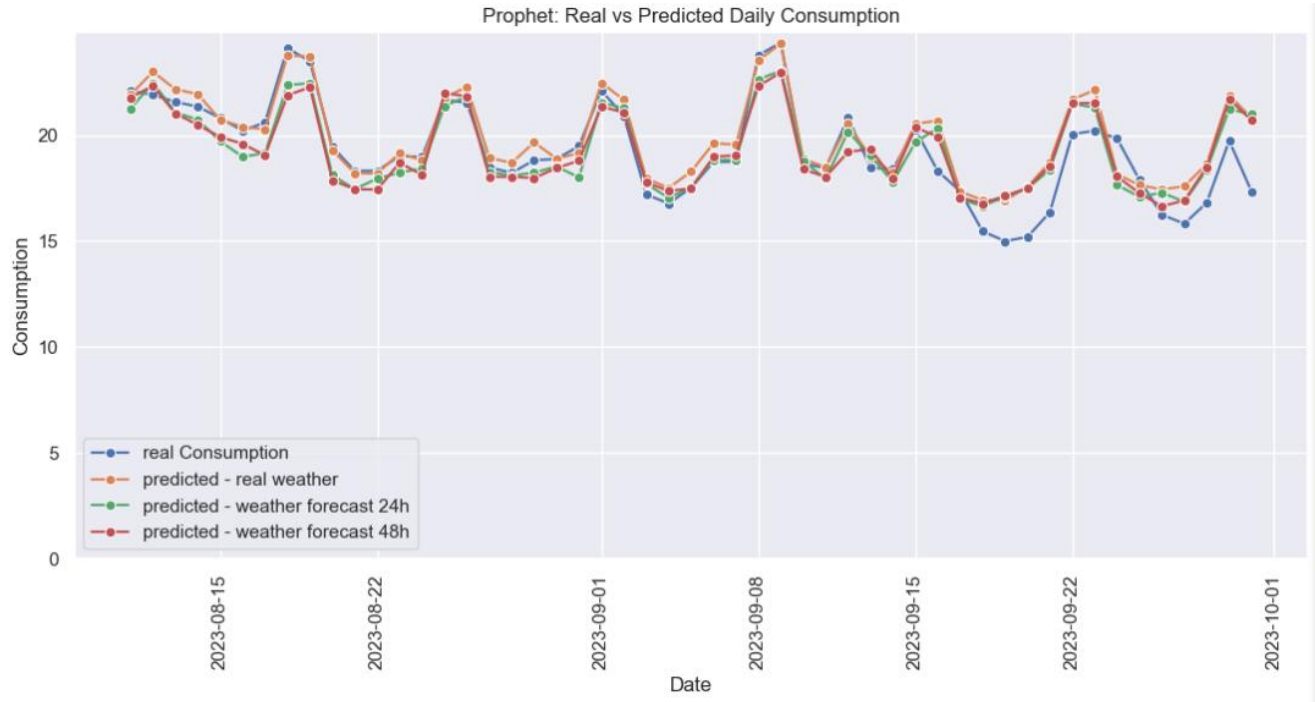


Figure 57: Comparison of real consumption with Obs, SD→SD, and PD→PD model predictions, Ashdod, 2023. Consumption figures are average daily kWh per meter (apartment).

To examine this relationship directly, Figure 58 and Figure 59 plot the hourly forecast error in the Heat Index (observed minus forecast) against the corresponding consumption prediction residual (observed minus predicted), averaged across buildings, for forecasts issued 24 and 48 hours ahead, respectively. Both figures show a clear positive relationship: when the Heat Index forecast underestimates the true value (positive x-axis), the consumption prediction also tends to underestimate the true consumption (positive y-axis), and vice versa. The correlation is similar in magnitude for both forecast horizons: $r = 0.4234$ for the 24-hour forecast and $r = 0.4242$ for the 48-hour forecast. This confirms that a meaningful share of the consumption prediction error when using forecasted weather is directly attributable to error in the underlying weather forecast.

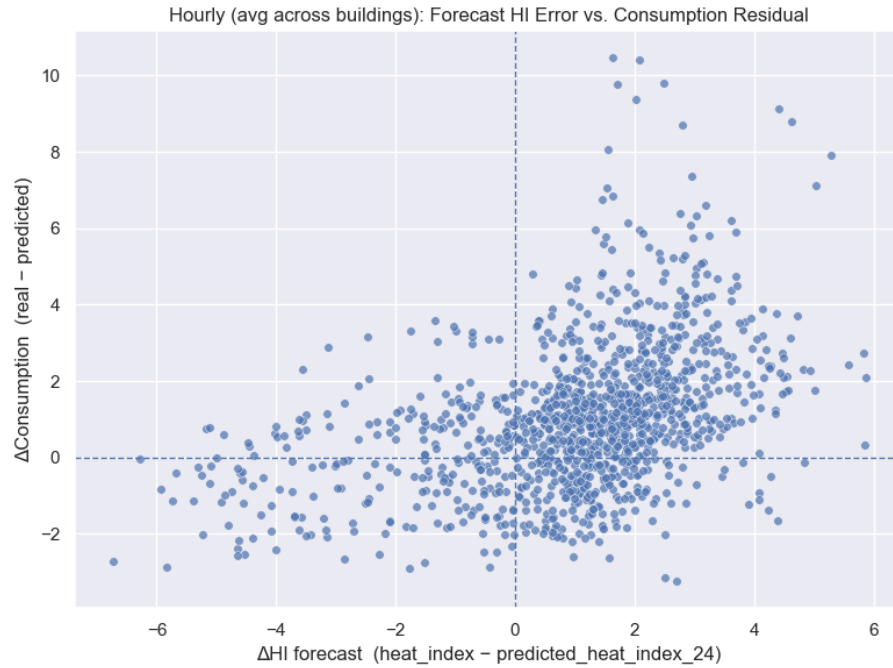


Figure 58: Hourly Heat Index forecast error vs. consumption prediction residual, 24-hour-ahead forecasts, Ashdod 2023 ($r = 0.4234$).

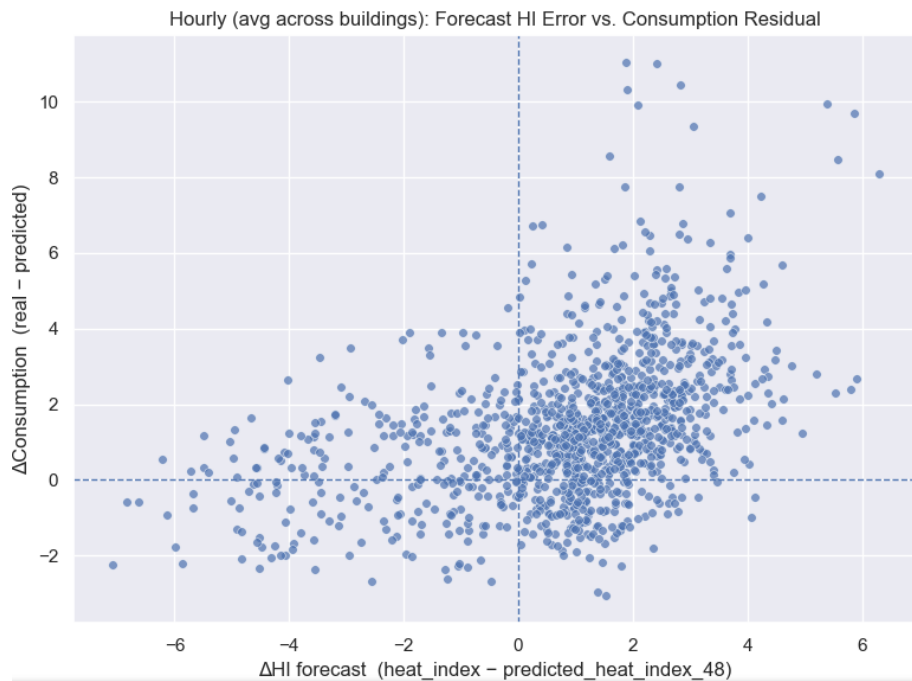


Figure 59: Hourly Heat Index forecast error vs. consumption prediction residual, 48-hour-ahead forecasts, Ashdod 2023 ($r = 0.4242$).

4. Conclusions and Future Work

In this part of the study, we developed a framework to predict electricity consumption in residential buildings across urban environments in Israel, integrating cross-building prediction, future forecasting, and feature-based classification and clustering. Key findings are as follows:

- LR and EBM models achieved high accuracy across all 7 cities, both at the city level (R^2 up to 0.995) and at the individual building level (R^2 up to 0.742).
- Physical building characteristics (floor area, number of floors, statistical zones) contributed negligibly to model accuracy, as their effect appears to be embedded in the historical consumption pattern of each building.
- The Prophet model produced accurate future forecasts at the city level (R^2 0.791–0.965, except for a single city and year) without relying on recent consumption data as input.
- Using weather forecasts instead of observed meteorological data caused only a marginal reduction in accuracy; models trained on forecasts outperformed models trained on observations when both were evaluated on forecast inputs.
- Per-building Prophet models aggregated to the city level consistently outperformed a single-series citywide model.
- Statistical feature extraction via tsfresh enabled city-of-origin and socioeconomic classification, as well as discovery of natural consumption-pattern clusters among buildings. K-Means clustering identified 7 distinct consumer types, with 3 large clusters accounting for roughly 75% of buildings; these consumption-based types only partly align with geographic and socioeconomic boundaries, and show moderate but incomplete persistence when buildings are re-clustered using data from a different time period.

Together, these findings show that researchers can build accurate, interpretable, and scalable electricity consumption prediction models for residential buildings using

accessible data, including historical consumption records and the Heat Index, without requiring complex building-level physical data.

The findings above come from a dataset that, while large, covers only two full years (2023–2024) and does not represent the Israeli residential building stock. We selected buildings based on smart-meter availability at the beginning of 2023, included only buildings with more than three residential units, and found them unevenly distributed within and across the seven cities. Consequently, the conclusions drawn here apply to the analysed buildings rather than to the cities or populations they belong to. As smart-meter installation expands nationwide, a larger, more representative dataset, stratified by climatic region, city size, building typology, and population group, and spanning several consecutive years, would allow us to address several open questions.

First, a multi-year record would let forecasting models incorporate yearly seasonality and holiday effects. It would also let forecasts extend beyond the summer season and a two-month horizon, and enable evaluation of forecast-based predictions over more than a single year of weather forecasts.

Second, it would provide a more reliable estimate of how consumption responds to extreme heat, since it would better represent heatwaves of different times of year, intensities, and durations. It would also allow researchers to distinguish long-term trends, such as changes in air-conditioning usage or the effects of rising temperatures, from year-specific fluctuations.

Third, repeating the clustering over several years would clarify which construction-consumption nexus types are persistent and which reflect transient conditions, given the only moderate cluster retention observed between 2023 and 2024.

Finally, linking a representative dataset to reliable building data (e.g., building age, construction technology, and apartment size) and to finer-grained demographic information would help disentangle the effect of construction from that of occupant demographics. In the present dataset, demographics appeared to be the dominant factor, as shown, for example, by the high consumption of buildings in low-socioeconomic, predominantly ultra-

Orthodox neighbourhoods. Such a dataset could also support predicting consumption for buildings with no consumption history, such as in planned neighbourhoods. Additionally, it could provide the empirical basis for validating and calibrating Urban Building Energy Models (UBEM) at the national scale, and for producing location- and population-specific estimates of future peak demand to support national energy planning.

Part II: Asymmetric approach to national-level prediction of electricity consumption

1. Introduction

The modern electricity grid operates through wholesale markets in which producers and large consumers interact - directly or through a central operator - to determine how much power is generated and consumed. In a stylised day-ahead market, producers submit offers specifying the quantities of electricity they are willing to supply at various price points, while suppliers (and possibly a handful of very large consumers) submit corresponding demand bids. A system operator then clears the system by finding the price at which aggregate supply equals aggregate demand, setting a single clearing price for each settlement period (typically one hour). This mechanism promotes economic efficiency by rewarding low-cost producers and incentivising demand response (Kirschen & Strbac, 2018; Stoft, 2002).

Real markets differ considerably in how centralised this process is. The Israeli market, which is the setting of this study, is comparatively centralised: the system operator (Noga) forecasts national demand for each hour of the following day, and this day-ahead forecast - together with the resulting price signal - determines how much electricity producers are scheduled to deliver in each settlement period. The process is therefore less the continuously cleared double auction of textbook descriptions and more a coordinated, forecast-driven dispatch. Regardless of the precise market design, however, the day-ahead demand forecast is the pivotal quantity: it is fixed before real-time demand is known, and its accuracy propagates directly into generation scheduling and price.

A central challenge in this architecture is that procurement decisions must be finalised a day in advance, before actual real-time demand is known. The accuracy of demand forecasts therefore directly affects the operational decisions and financial outcomes of all market participants. It remains an open question to what extent individual suppliers maintain their own demand models. A plausible account of current practice is that, once the operator publishes its day-ahead demand forecast, each supplier simply draws its own share from that forecast based on experience, and the market settles on a relatively stable price that is reasonable on average but may embed sizeable, systematic error margins. If so,

more sophisticated forecasting - whether by the suppliers or by the system operator - could translate into lower retail prices, which is part of what motivates the present work.

A system operator that accurately anticipates demand can procure the right amount of energy at competitive prices. Inaccurate forecasts, by contrast, force participants into costly corrective actions in the balancing or real-time markets, where prices can be dramatically different from day-ahead levels. It is worth stressing that these price differences arise from two distinct sources of uncertainty: errors in predicting *demand*, and errors in predicting available *production capacity* - the latter being especially pronounced in systems that rely heavily on variable renewable generation such as solar and wind (Morales et al., 2014). This paper is concerned exclusively with the first source of differences, uncertainty in day-ahead demand; production-side uncertainty, while important, is outside our scope.

1.1 Asymmetric costs in electricity markets

A critical and often overlooked aspect of electricity demand forecasting is the *asymmetry* of the costs associated with forecast errors. When demand is *overestimated*, the system operator or retailer procures more electricity than is ultimately needed. The excess supply can typically be sold back into the balancing market, curtailed, or stored (if storage is available), resulting in a relatively modest financial loss. In deregulated markets with well-functioning balancing mechanisms, the per-unit cost of over-procurement is generally bounded and predictable.

When demand is *underestimated*, the cost of error is usually much higher. The system operator must source additional electricity in the real-time market at short notice to prevent blackouts. Real-time prices are typically highly volatile and can spike to several multiples of the day-ahead price during periods of scarcity. In extreme cases, the system may lack sufficient reserve capacity to meet demand, triggering involuntary load shedding, emergency imports, or the activation of expensive peaking units. Beyond the direct financial costs, supply shortfalls carry significant reliability and regulatory implications (Kirschen & Strbac, 2018; Stoft, 2002).

Empirical studies in various electricity markets suggest that the ratio of under-procurement costs to over-procurement costs, c_u/c_o , typically lies in the range [5, 20], depending on market structure, regulatory context, and the availability of real-time resources (Zhang et al., 2022). This asymmetry fundamentally changes the optimal forecasting problem: rather than seeking a prediction that minimises expected squared error, a rational market participant should seek a prediction that minimises expected asymmetric cost.

This decision-theoretic view has developed alongside a substantial body of work on electricity forecasting more broadly. Weron (2014) provides a comprehensive review of the state of the art in electricity price forecasting, surveying statistical, machine-learning, and hybrid approaches. While its focus is on price rather than demand, the review highlights the central role of model selection, loss-function choice, and benchmark comparisons that also apply in the load-forecasting context.

Hong and Fan (2016) offer a tutorial review of probabilistic electric load forecasting, covering methods ranging from regression-based quantile estimation to ensemble and Bayesian approaches. Crucially, they argue that point forecasts are insufficient for operational decision-making and advocate for full predictive distributions or at least prediction intervals - a view consistent with the quantile-forecasting objective we adopt.

The practical importance of probabilistic evaluation metrics, and of the pinball loss in particular, was cemented by the Global Energy Forecasting Competition 2014 (Hong et al., 2016). GEFCom2014 was the first major forecasting competition to use pinball loss as its primary scoring rule across the electricity load, price, wind, and solar tracks. It established pinball loss as a community standard and demonstrated that models explicitly optimised for quantile accuracy substantially outperform those adjusted only post-hoc.

Despite the theoretical clarity of this framing, most practical forecasting systems in electricity markets continue to use symmetric loss functions - most commonly mean squared error (MSE) or mean absolute error (MAE) - both for training predictive models and for evaluating their performance. This disconnect between training objective and

operational cost constitutes a systematic source of suboptimality that has received growing attention in the academic literature (Lin et al., 2020; Wang et al., 2019; Wu et al., 2021; Zhang et al., 2022) but has yet to see widespread adoption in practice.

1.2 Cost-oriented and asymmetric learning for load forecasting

The theoretical foundation for prediction under asymmetric costs was laid by Koenker and Bassett (1978), who introduced *regression quantiles* as a natural generalisation of sample quantiles to the linear model. They showed that minimising the asymmetric pinball loss (equivalently, the “check function”) yields an estimator of the conditional τ -quantile of the response, and established its asymptotic properties. This result directly underpins our use of a cost-sensitive asymmetric loss function - instantiated as a non-linear variant of the pinball loss - as the training objective for day-ahead demand forecasting.

Christoffersen and Diebold (2009) provided the corresponding decision-theoretic foundation. They showed that under a general asymmetric loss function, the optimal point forecast is *biased*, and that the optimal amount of bias is time-varying and depends on higher-order conditional moments such as conditional variance. In particular, their analysis implies that any forecast trained with a symmetric objective (e.g. MSE) will be systematically suboptimal when the true cost structure is asymmetric, even if the conditional mean is estimated perfectly.

Building on these foundations, a more recent line of work addresses the mismatch between symmetric training objectives and asymmetric operational costs directly in the electricity setting, embedding the cost asymmetry into the learning objective itself rather than correcting for it after the fact.

Wang et al. (2019) trained a Long Short-Term Memory (LSTM) network directly with the pinball loss to produce probabilistic individual load forecasts. By guiding the network to minimise $\mathbb{E}[\mathcal{L}_\tau]$ rather than MSE, their approach yields calibrated quantile predictions without requiring post-hoc quantile regression on the residuals. Their results show improved reliability and sharpness compared to symmetric-loss baselines.

Lin et al. (2020) addressed the related problem of contract-capacity optimisation for high-voltage consumers, where the electricity pricing structure is inherently asymmetric: underestimating demand triggers penalty charges at a higher per-unit rate than overestimating. They proposed several asymmetric loss functions derived directly from the Taiwan Power Company's pricing formula, and demonstrate that LSTM models trained with these loss functions reduce electricity costs by 0.88%–2.42% over symmetric MSE baselines.

Wu et al. (2021) extended this idea to support vector regression (SVR), developing a cost-oriented asymmetric SVR (AsySVR) framework based on a linear-linear loss function with an insensitivity parameter to improve generalisation. Evaluated on New South Wales load data, their asymmetric framework achieved daily economic cost reductions of 42%–57% relative to standard SVR, underscoring that the choice of loss function can have a large practical impact on operational outcomes.

Zhang et al. (2022) further formalised the decision-oriented perspective on load forecasting, framing the problem explicitly in terms of the downstream cost incurred by the system operator. Their cost-oriented framework bridges the gap between forecast-accuracy metrics and operational objectives, and provides theoretical justification for replacing standard point-forecast losses with cost-aware alternatives.

Taken together, this recent body of work establishes that asymmetric, cost-sensitive training objectives are both theoretically motivated and empirically superior to symmetric alternatives in electricity applications. Two gaps, however, remain. First, existing studies almost exclusively employ *piecewise-linear* asymmetric losses - the pinball loss, or the linear-linear penalties derived from a specific tariff - and thus inherit the newsvendor model's linearity assumption, even though real procurement costs are frequently non-linear. Second, and more practically, these methods are typically benchmarked against statistical baselines rather than against the forecast an experienced system operator actually deploys, leaving open the question of whether cost-aware training improves upon a strong, real-world operational forecast.

1.3 Contribution

Our work is designed to overcome both gaps currently existing in the literature described above. We formulate the forecasting problem with a general, not necessarily linear asymmetric cost function that subsumes the pinball loss as a special case, and train a neural network to minimise it end-to-end. Crucially, we evaluate against the *real* day-ahead forecasts published by the Israeli system operator Noga - an operational baseline with access to proprietary data and domain expertise - rather than against a synthetic statistical model. This lets us ask not merely whether cost-aware training beats a symmetric loss in the abstract, but whether it can improve upon the forecast currently used to help run a national grid. The contributions of this paper are thus as follows:

1. **A general asymmetric-cost formulation.** We formulate day-ahead demand forecasting under a general, not-necessarily-linear asymmetric cost function that subsumes the pinball loss as a special case, and we train a feedforward neural network to minimise it end-to-end.
2. **Evaluation against a real operational forecast.** Rather than benchmarking against a synthetic statistical baseline, we compare against the *actual* day-ahead forecasts published by the Israeli system operator Noga - a strong, real-world baseline produced with proprietary data and domain expertise.
3. **A practical model that improves on the operator’s forecast.** Although raw predictive accuracy was *not* a goal of this work, our model outperforms the Noga forecast not only on cost-sensitive metrics but even under symmetric error metrics such as mean squared error. We regard this as an incidental by-product of a deliberately simple architecture and believe that substantial further improvement is readily achievable.
4. **A demonstration of economic impact.** We quantify the operational cost reduction that cost-aware training delivers over symmetric baselines across a range of cost scenarios, showing an effect that is large enough -and cheap enough to obtain - to be practical to adopt in a live market setting.

5. **An open-source implementation.** We release our complete code and data pipeline, from raw data ingestion through model training and evaluation, as open source,¹ so that our results can be reproduced and extended.

2. Cost-sensitive forecasting: Problem formulation and asymmetric loss

2.1 Problem formulation

In day-ahead electricity demand forecasting, the system operator must commit to a procurement quantity p before observing actual demand X . We model X as a random variable whose distribution is *unknown* and must be inferred from data; we make no parametric assumption about its form.

2.1.1 Assumptions

We adopt the following assumptions, which are mild and operationally grounded:

1. **Asymmetry.** Under-prediction is more costly than over-prediction of equal magnitude: the costs incurred when actual demand exceeds the forecast are larger than those incurred when the forecast exceeds actual demand ($c_u > c_o > 0$).
2. **Monotonicity.** Costs are non-decreasing in the magnitude of the prediction error: larger errors are never cheaper than smaller errors, in either direction.
3. **Known, differentiable costs.** For the purposes of this paper we assume that the cost functions are known and differentiable, so they can be incorporated directly as training objectives for gradient-based optimisers.

¹ <https://github.com/gkutiel/noga>

2.1.2 Cost structure

Let $\ell_o: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ and $\ell_u: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ denote the cost functions for over- and under-prediction, respectively, where the argument is the magnitude of the error. The total realised cost of a procurement decision p against actual demand X is:

$$C(p, X) = \begin{cases} \ell_o(p - X) & \text{if } X \leq p \quad (\text{over-prediction}) \\ \ell_u(X - p) & \text{if } X > p \quad (\text{under-prediction}) \end{cases}$$

2.1.3 Goal

The objective is to find a procurement decision p^* that minimises the *expected cost*, and not the expected prediction error.

2.1.4 The newsvendor special case

To gain analytical insight, we first consider the classical special case in which costs are *linear*, the setting of the *newsvendor model* (Arrow et al., 1951), a classical inventory-theoretic framework in which a decision-maker commits to a quantity before demand is realised. Let $c_o > 0$ and $c_u > c_o$ be the per-unit overage and underage costs. This is a well-known problem in which the optimal procurement level is given by the critical fractile formula:

$$p^* = F^{-1}\left(\frac{c_u}{c_u + c_o}\right)$$

2.2 Limitations of the linear newsvendor framework

While the newsvendor model provides a tractable and theoretically grounded formulation, it does not always capture the complexities of real-world electricity markets. Our general formulation (Equation [eq:cost]) addresses this limitation: it does not assume linear costs and it does not require knowledge of the demand distribution, relying instead on empirical samples via the training loss.

2.3 Synthetic cost functions for evaluation

Since we do not have access to actual cost data from the electricity market operator, we evaluated our models against *synthetic* cost functions that satisfy the qualitative properties

stated in the problem formulation. For simplicity, we conducted our main evaluation using two *pinball loss* functions with under/over-prediction cost ratios of 1:5 and 1:20, respectively. These ratios are consistent with the findings of prior work on asymmetric costs in electricity demand forecasting, which reports that the cost of under-prediction typically exceeds the cost of over-prediction by a factor of five to twenty (Weron, 2014). Pinball losses are piecewise linear and widely used in quantile regression, making them a natural and interpretable baseline for cost-sensitive evaluation.

3. Methods

The objective of this study was to determine whether a demand forecaster trained end-to-end to minimise an *asymmetric operational cost* can outperform both a symmetric-loss model and the real operational forecast of the Israeli system operator, under the cost asymmetries characteristic of electricity procurement. This section describes the material and procedures used to answer that question: the dataset (Section 3.1), the neural architecture and the three training objectives we compare, and the protocol under which they are evaluated (Section 3.2).

3.1 Dataset

The dataset is collected from the Israeli electricity market. Israel is a small country in the Middle East, covering about 20,770 km² and home to roughly 10 million people as of 2026. The population geographic distribution is highly concentrated: the vast majority of Israelis live along the narrow Mediterranean coastal plain, a strip approximately 150 km long stretching from Haifa in the north through the greater Tel Aviv metropolitan area and on to Ashdod in the south. Figure 1 illustrates this concentration and directly motivates our choice of weather stations: Tel Aviv, Jerusalem, and Haifa are the three largest urban centres and together capture the climatic conditions experienced by the bulk of the population, whose electricity use drives the national demand signal.

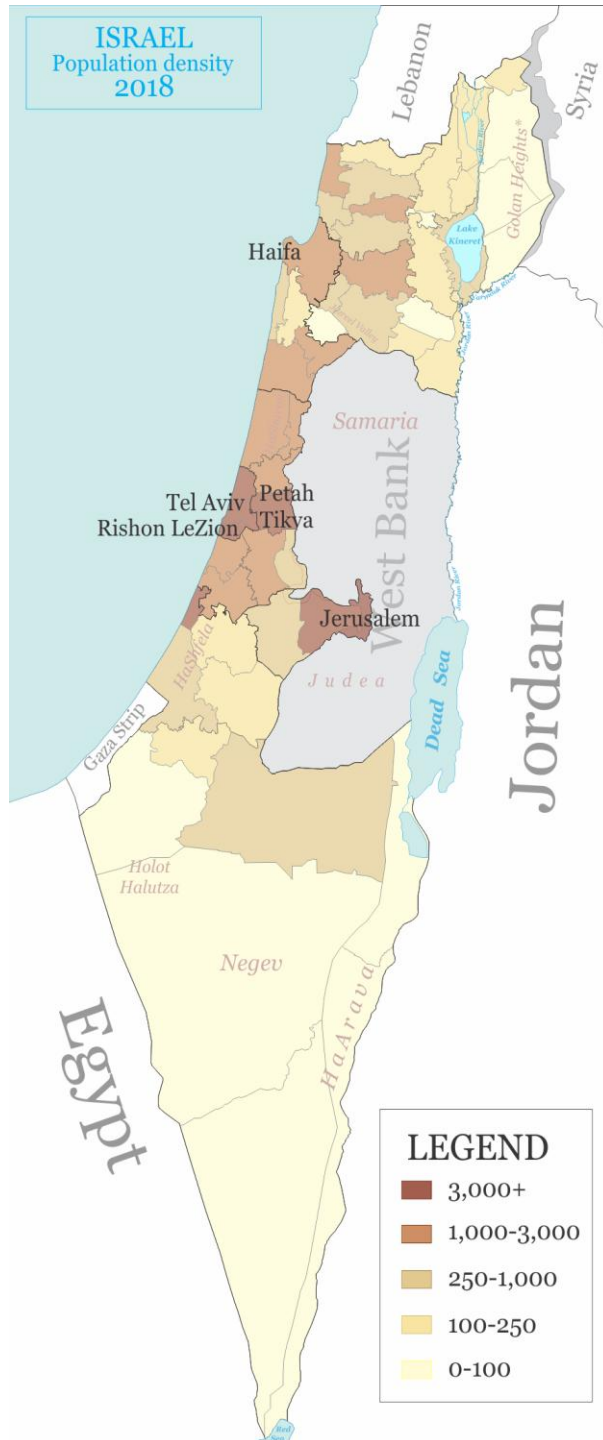


Figure 1: Population density of Israel (2018). The coastal plain, running approximately 150 km from Haifa to Ashdod, concentrates the overwhelming majority of the population. This spatial clustering justifies using weather observations from Tel Aviv, Haifa, and Jerusalem as the primary climatic features for demand forecasting. Population density is per 1 square kilometre (source: Yhockey, Wikimedia, https://en.wikipedia.org/wiki/Demographics_of_Israel#/media/File:Israel_population_density_2018.png).

The dataset consists of two complementary sources: electricity consumption data obtained from the Israeli electricity system operator Noga,² and weather data sourced from the Israeli Meteorological Service (IMS).³ Electricity consumption data includes readings of power usage at 5-minute intervals. Crucially, it also includes the *real day-ahead forecasts* published by Noga as part of the market clearing process. These are the actual forecasts used by the system operator in the day-ahead market, not post-hoc reconstructions. Their inclusion in our dataset provides two important opportunities: (i) we can directly compare the performance of our learned models against the operational forecasts actually used in the market, serving as a strong real-world baseline; and (ii) we can analyse the distribution of real-world prediction errors produced by an experienced operator, which motivates and contextualises the cost asymmetries studied in this paper.

The weather data include air temperature, wet-bulb temperature, and relative humidity, whose combined effect drives building cooling loads, collected every three hours in Israel's three major cities: Tel Aviv, Jerusalem, and Haifa. In Israel, where the percentage of households with air conditioning systems is currently 96.5% (ranging from 89% in the lower decile to 98.3% in the upper decile) according to data from Israel's Central Bureau of Statistics, there is a strong correlation between electricity consumption levels and building cooling and heating loads resulting from changing weather conditions and building occupancy trends.

3.1.1 Data analysis

To better understand the dataset, we performed an exploratory data analysis (EDA) to identify patterns and relationships between the electricity consumption and weather data. Figure 2 shows daily electricity consumption over the dataset period (1 March 2023 to 28 February 2026) alongside the daily average air temperature in Tel Aviv.

² <https://www.noga-iso.co.il/statistics/annual-hourlyconsumption-data/>

³ https://ims.gov.il/he/data_gov

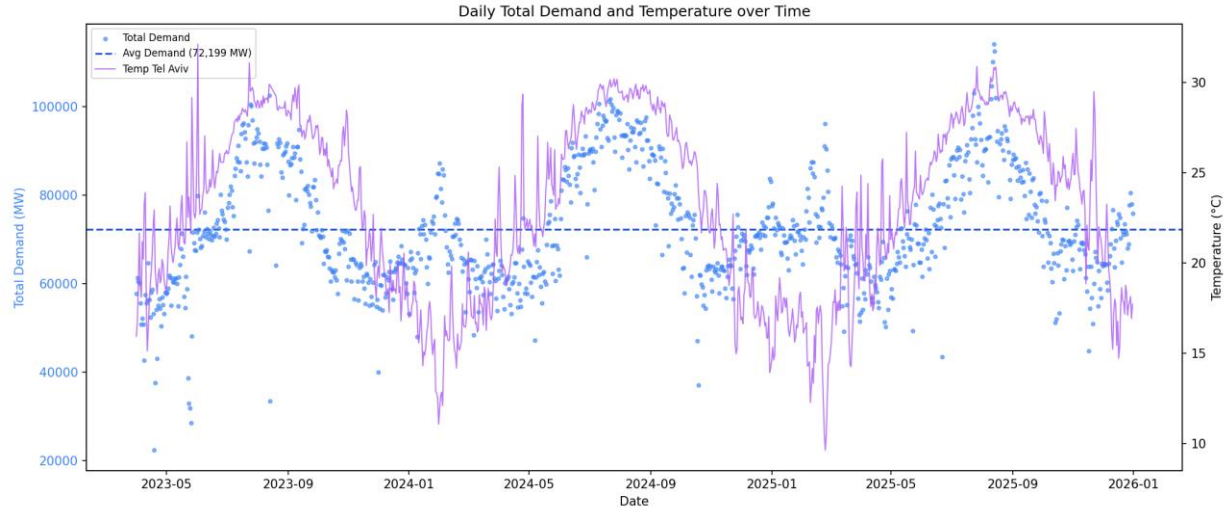


Figure 2: Daily electricity consumption (MWh) and observed air temperature over the dataset period (1 March 2023 to 28 February 2026). Clear seasonal patterns are visible, with a summer peak (July–August) driven by air-conditioning loads and a secondary winter peak (January–February) driven by heating.

Figure 3 shows the relationship between daily electricity consumption and the daily average air temperature in Tel Aviv.

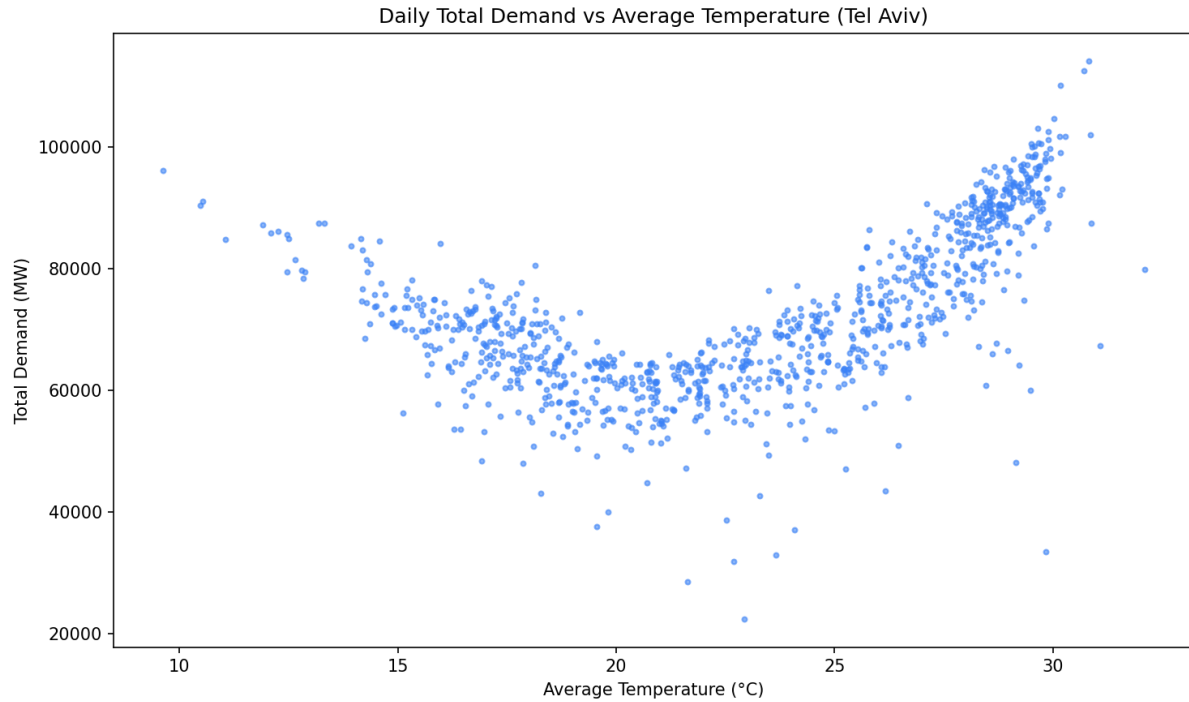


Figure 3: Daily electricity consumption vs. daily mean air temperature in Tel Aviv. The characteristic U-shaped relationship is visible, with high demand at both high air temperatures (cooling loads) and low air temperatures (heating loads). The balance-point air temperature is approximately 20 °C.

Figure 4 shows the KDE histogram of the forecast error, i.e., the difference between the actual demand and the day-ahead forecast.

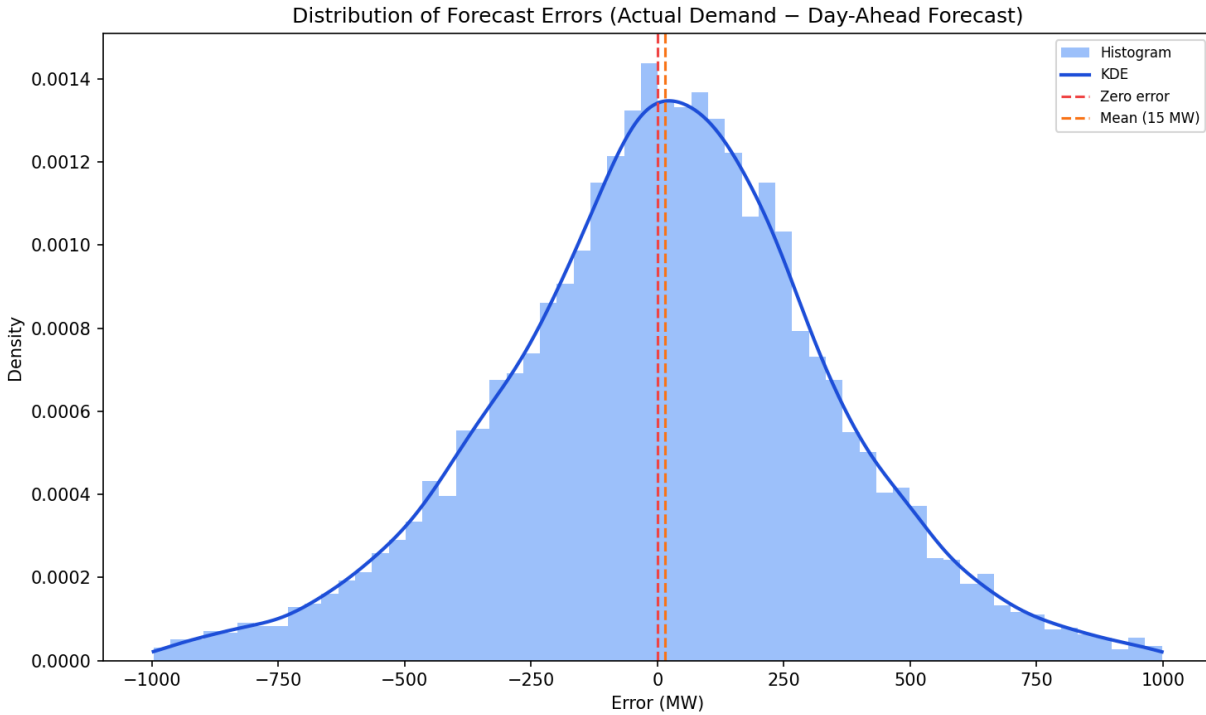


Figure 4: KDE histogram of the forecast error (actual demand minus day-ahead forecast). The distribution is approximately zero-mean with slightly heavier right tails, suggesting that large under-forecasts are marginally more frequent than large over-forecasts.

Table 1 reports summary statistics for the key variables in the daily dataset. The data span 1,096 days from 1 March 2023 to 28 February 2026.

Table 1: Summary statistics for daily variables ($N = 1,096$ days, 1 March 2023 to 28 February 2026).***Three anomalous forecast values excluded from statistics; see text.**

Variable	Mean	Std	Min	Median	Max
Total demand (MWh)	73,421	14,852	19,742	70,435	114,103
Day-ahead forecast (MWh)	73,109	14,680	2,400	70,200	116,148
Forecast error (MWh)	+312	3,241	-12,300	+284	+11,543
Air temp. Tel Aviv (°C)	22.1	4.8	9.6	22.5	30.3
Air temp. Jerusalem (°C)	18.9	6.4	2.3	19.8	36.5
Air temp. Haifa (°C)	20.3	5.1	4.4	20.8	31.3

3.1.2 Forecast error distribution

As shown in Figure 4, the Noga day-ahead forecast errors (actual minus forecast) are strikingly close to zero-mean and nearly symmetric: the mean error is only +312 MWh against a standard deviation of 3,241 MWh, and the KDE is approximately bell-shaped with modest skewness. In other words, the operational forecast is essentially unbiased, under-predicting and over-predicting with roughly equal frequency and magnitude.

From a pure accuracy standpoint, this is desirable. From a cost-sensitive standpoint, however, it is a warning sign. As established in Section 2, when the costs of under-prediction substantially exceed those of over-prediction, the cost-minimising forecast is *not* the conditional mean but a higher quantile - a deliberate upward bias proportional to the cost asymmetry. An approximately symmetric, zero-mean forecast therefore fails to account for the market's asymmetric cost structure and will systematically incur higher expected operational costs than a cost-aware alternative. This observation motivates our approach and is quantified empirically in Section 4.

Figure 5 breaks down the Noga forecast error magnitude by hour of the day and day of the week. Each cell shows the mean *percentage* absolute error ($|actual - forecast|/actual$) for that hour-of-week slot over the entire dataset. The heatmap reveals that forecast accuracy varies substantially across the week: certain hour-of-day and day-of-week combinations are consistently harder to forecast than others, with notable differences between weekdays, Friday, and Saturday. This structured, non-uniform pattern has an

important implication for post-hoc calibration: a global affine correction cannot account for the fact that errors are large in some slots and small in others. We return to this point in Section 4.

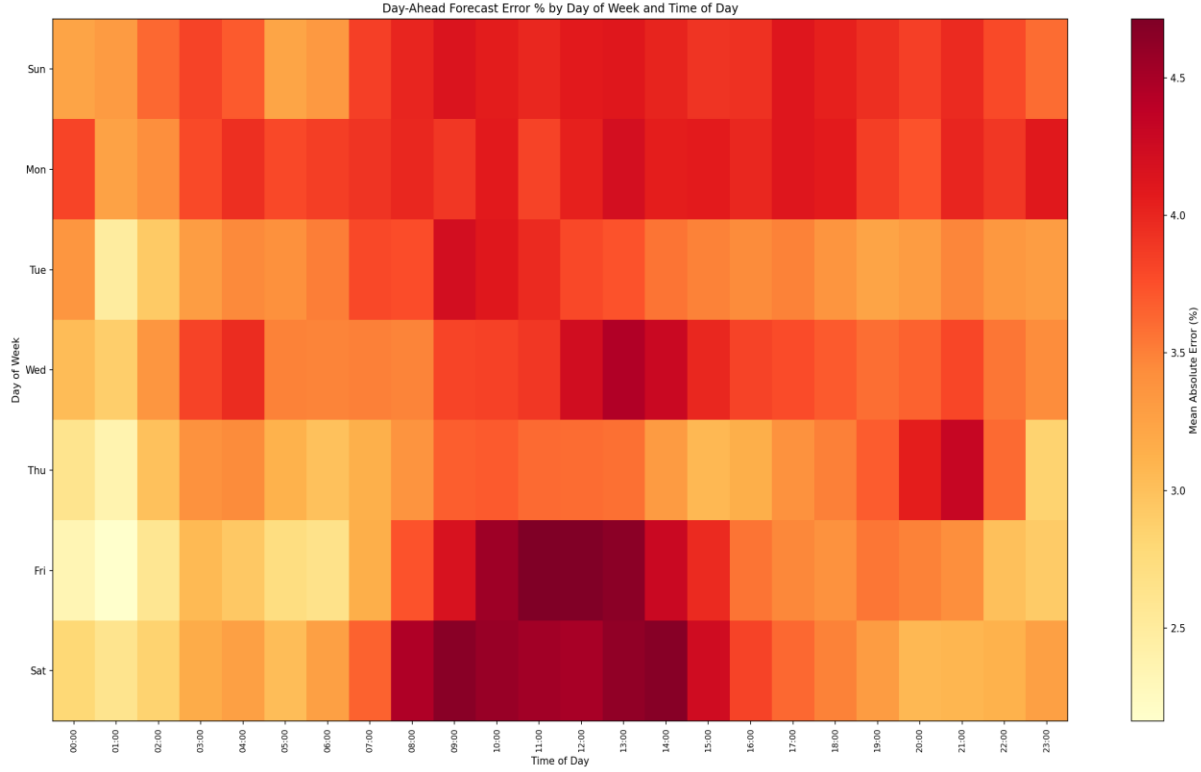


Figure 5: Heatmap of mean percentage absolute error ($|actual - forecast|/actual$) for Noga day-ahead forecasts, broken down by hour of the day (rows) and day of the week (columns). Each cell represents a specific hour-of-week slot. The non-uniform pattern demonstrates that forecast difficulty varies systematically across hours and days, and cannot be corrected by a single global offset.

3.1.3 Comparison with the Noga operational forecast

A distinctive feature of our dataset is the availability of the *real* day-ahead forecasts produced by Noga and used in the actual electricity market. This allows us to evaluate our learned models not merely against simple statistical baselines but against the operational forecast of an experienced system operator with access to proprietary data and domain expertise.

The Noga forecast achieves $R^2 \approx 0.98$ against actual demand, confirming it is a strong accuracy baseline. Its error distribution is, as discussed in Section 3.1.2, nearly zero-mean and approximately symmetric. While this is impressive from an accuracy standpoint, it also

reveals a limitation: the Noga forecast behaves as a conditional-mean estimator, making no explicit adjustment for cost asymmetry. A symmetric, unbiased forecast minimises mean squared error but, as we show empirically in Section 4, it does *not* minimise expected operational cost when under-prediction is substantially more costly than over-prediction. Its high R^2 is therefore a misleading indicator of operational performance under asymmetric costs - a point that motivates the cost-sensitive approach developed in this paper.

3.2 Model Training

All model variants share the same architecture: the prediction is expressed as a *soft attention* over a short window of recent historical demand values plus a learned additive correction. Weather conditions (air temperature, wet-bulb temperature, and relative humidity, as reported by the IMS, at three meteorological stations) and temporal patterns (time of day, day of week, month of year) jointly determine both the attention weights and the correction term. Temporal features are encoded via learned embeddings; weather features enter as continuous inputs. The model operates at 5-minute granularity and produces a day-ahead forecast: given conditions at a target time t , it predicts demand exactly 24 hours earlier in the input window - making it a suitable testbed for isolating the effect of the training objective on cost-aware forecasting. The variants differ *only* in their training objective, so that any difference in cost performance can be attributed to that objective alone.

3.2.1 Data splitting

The dataset is partitioned chronologically at 1 January 2025:

- **Training set:** March 2023 – December 2024 (≈ 672 days).
- **Test set:** January 2025 – February 2026 (≈ 424 days), held out for final evaluation.

The chronological split ensures no future information leaks into training.

3.2.2 Neural network architecture

For each target 5-minute slot t , the model receives three types of input:

- **Historical demand.** A window of $H = 10$ consecutive 5-minute demand readings $\mathbf{h} = (h_1, \dots, h_H)$ ending exactly 24 hours before t (i.e. the same time-of-day window from the previous day).
- **Weather features.** Nine continuous variables measured at time t : air temperature, wet-bulb temperature, and relative humidity for three stations (Haifa, Jerusalem, Tel Aviv).
- **Temporal embeddings.** Learned embeddings for time of day (288 positions, dimension 16), day of week (7 positions, dimension 3), and month of year (12 positions, dimension 2).

These are concatenated into a vector of dimension $H + 9 + 16 + 3 + 2 = 40$ and passed through a small MLP with two hidden layers of width 32 and LeakyReLU activations,⁴ producing an output vector $\mathbf{o} \in \mathbb{R}^{H+1}$. The first H outputs are treated as attention logits over the history window; the last output is an additive correction scalar δ . The forecast is:

$$\hat{y} = \sum_{k=1}^H \text{softmax}(\mathbf{o}_{1:H})_k h_k + \delta,$$

where all demand values are normalised by a constant scale factor $s = 100$ during training and rescaled at inference time. The model thus learns to select the most informative historical reading given current conditions and applies a weather- and calendar-conditioned correction. Both models (L1 and asymmetric) are trained with the Adam optimiser (Kingma & Ba, 2015) at a learning rate 2×10^{-4} , batch size 10 000, and up to 500 epochs with early stopping (patience 20, evaluated every 10 epochs).

3.2.3 Training with L1 Loss

The symmetric neural network baseline is trained by minimising the mean absolute error (L1 loss):

⁴ <https://docs.pytorch.org/docs/2.13/generated/torch.nn.LeakyReLU.html>

$$\mathcal{L}_{L1}(\hat{y}, y) = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|$$

L1 loss treats over- and under-prediction symmetrically, and yields predictions that converge to the conditional median. As discussed in Section 2, this is suboptimal when costs are asymmetric.

3.2.4 Post-hoc cost-optimised baseline

To further isolate the benefit of end-to-end cost-aware training, we include an additional baseline in which predictions from the L1-trained neural network are adjusted *after* training by a learned affine transformation that minimises the expected operational cost on the training set. Concretely, we train a linear calibration model $g(z) = \alpha z + \beta$ by minimising:

$$(\alpha^*, \beta^*) = \operatorname{argmin}_{\alpha, \beta} \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} \mathcal{L}_{\text{asym}}(\alpha \hat{y}_i + \beta, y_i)$$

where $\hat{y}_i$ are the L1-trained predictions and $\mathcal{L}_{\text{asym}}$ is the asymmetric loss (Section 2). The calibration is then applied uniformly to all test-set predictions.

This baseline tests whether a practitioner could recover most of the cost reduction by post-hoc rescaling a symmetric model, without modifying the training objective. Its key limitation is that it can only apply a global affine correction: it cannot adapt the *shape* of the predictive response to local input features, nor exploit the asymmetric cost structure during back-propagation. Cost-aware training, by contrast, reshapes the model’s response to every input pattern throughout optimisation, enabling feature-adaptive upward adjustments that are largest precisely when under-prediction risk is highest.

3.2.5 Training with asymmetric loss

The cost-aware neural network is trained by minimising the asymmetric cost function (Section 2) directly as the training objective. Because the loss is differentiable, it is incorporated into the standard gradient-based training loop with no additional computational overhead compared to L1 training.

3.2.6 Evaluation protocol

Using the architecture and training loop above, we compared three training strategies, differing only in their objective:

- **L1 (MAE)** - trained with mean absolute error (symmetric baseline).
- **Pinball 5:1** - trained with the pinball loss at a cost ratio of 5:1 (under-prediction penalised 5× more than over-prediction).
- **Pinball 20:1** - trained with the pinball loss at a cost ratio of 20:1.

To isolate the benefit of end-to-end cost-aware training, we additionally evaluated *post-hoc calibrated* variants of all three models (Section 3.2.4), in which an affine transformation $g(z) = \alpha z + \beta$ is fitted on the training set to minimise the expected asymmetric cost.

We evaluated the models on three cost functions that reflect different assumptions about the cost of over- and under-prediction: the L1 cost, the Pinball 5:1 cost, and the Pinball 20:1 cost. The primary metrics are the **Pinball 5:1** and **Pinball 20:1** costs, which best reflect the asymmetric structure of real-world electricity procurement costs. Figure 6 illustrates these three cost functions: the Pinball 5:1 and Pinball 20:1 functions penalise under-predictions more heavily than over-predictions, reflecting the real-world asymmetry of electricity procurement costs, while L1 serves as a symmetric reference point.

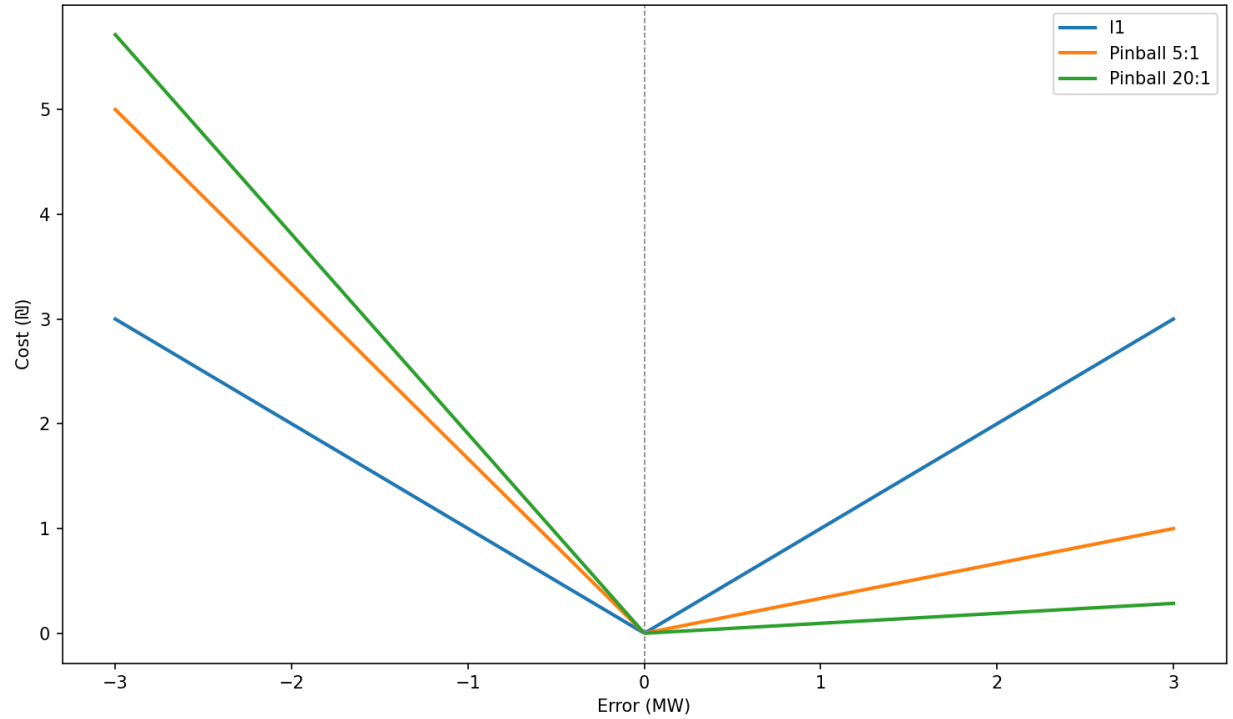


Figure 6: The three cost functions used for evaluation: L1 (MAE, symmetric baseline), Pinball 5:1, and Pinball 20:1. The pinball functions penalise under-predictions more heavily than over-predictions, with the 20:1 variant applying a much steeper penalty on the under-prediction side.

4. Results and discussion

Table 2 reports the average test-set cost of each model under each evaluation cost function, before and after post-hoc calibration.

Table 2: Test-set cost under three evaluation metrics. Lower is better. Bold marks the best value in each column. *Top block*: raw predictions; *bottom block*: after post-hoc affine calibration.

Model	L1	Pinball 5:1	Pinball 20:1
<i>Without post-hoc calibration</i>			
L1	2.704	3.045	3.167
Pinball 5:1	3.652	1.820	1.166
Pinball 20:1	6.175	2.245	0.841
<i>With post-hoc affine calibration</i>			
L1 + calibration	2.704	1.979	1.022
Pinball 5:1 + calibration	2.915	1.820	0.851
Pinball 20:1 + calibration	3.424	1.964	0.841

Figure 7 visualises the same results as a colour-coded model $\times$ loss matrix, making it easy to see at a glance which model wins on each metric and how calibration shifts the ranking.

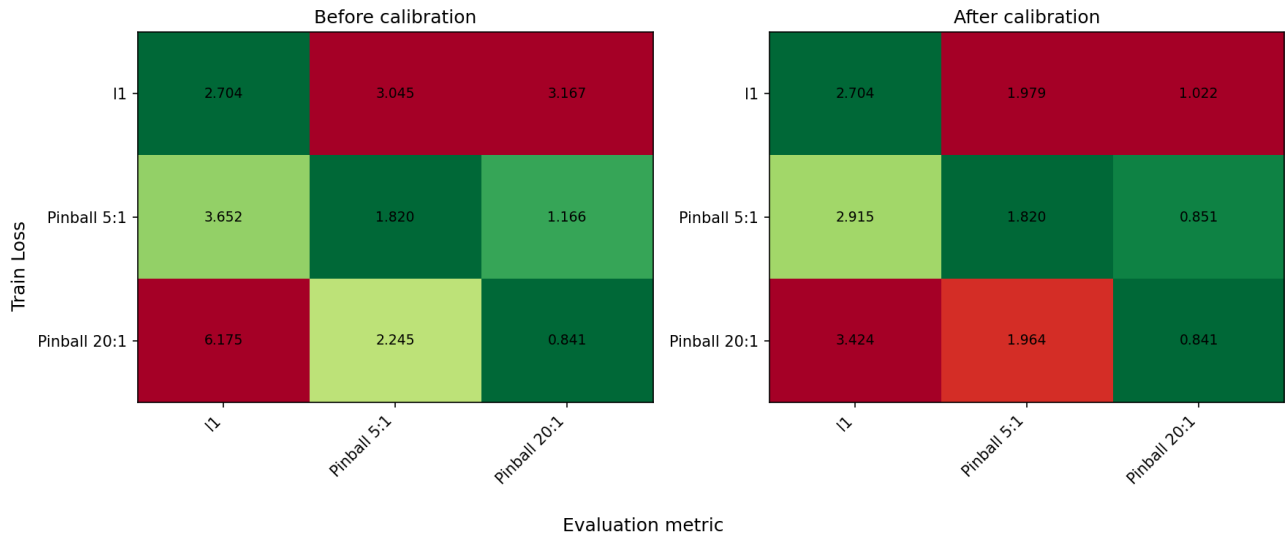


Figure 7: Train-loss $\times$ evaluation-metric matrix for the three models (L1, Pinball 5:1, Pinball 20:1), before calibration (left) and after post-hoc affine calibration (right). Each cell shows the test-set cost; green marks the best value in each column and red the worst. The diagonal tendency - each model achieving its lowest cost on the metric it was trained with - is clearly visible. Calibration reduces off-diagonal costs substantially (e.g. L1 model under Pinball 20:1 drops from 3.167 to 1.022) while leaving the diagonal unchanged.

4.1 Each model excels on its own training metric

Without calibration, Figure 7 shows clear diagonal dominance: each model achieves its lowest cost on the metric it was trained to minimise. The L1 model produces the lowest L1 cost (2.704), the Pinball 5:1 model the lowest Pinball 5:1 cost (1.820), and the Pinball 20:1 model the lowest Pinball 20:1 cost 0.841. This confirms that the training objective directly shapes the resulting forecast's cost profile.

4.2 Asymmetric training substantially outperforms the symmetric baseline

The advantage of cost-aware training grows with the asymmetry ratio. Under the Pinball 5:1 metric, the Pinball 5:1 model (1.820) outperforms the L1 baseline (3.045) by 40%. Under the more demanding Pinball 20:1 metric, the Pinball 20:1 model (0.841) outperforms L1 (3.167) by 73%. Figure 8 shows the mechanism: the Pinball 20:1 model produces a heavily right-

skewed error distribution, deliberately over-predicting to avoid costly under-procurement events, while the L1 model is symmetric and centred near zero.

4.3 Post-hoc calibration narrows but does not close the gap

Affine calibration substantially improves the L1 model on asymmetric metrics (Pinball 20:1: $3.167 \rightarrow 1.022$; Pinball 5:1: $3.045 \rightarrow 1.979$), confirming that a global bias correction captures much of the asymmetric signal. However, calibrated L1 still falls short of the dedicated pinball models (Pinball 5:1: 0.851; Pinball 20:1: 0.841), because a single affine shift cannot replicate the input-conditional upward bias that end-to-end cost-aware training achieves. Figure 5 (Section 3.1.2) illustrates exactly why: forecast error magnitude varies substantially by hour of the day and day of the week. A uniform global correction applies the same rescaling to every prediction regardless of context; only a model that conditions its output on time features can adapt its bias to each slot individually.

4.4 Error distributions reflect the targeted quantile

Figure 8 shows the prediction error distributions on the test set for each model. The L1 model is symmetric and centred near zero, as expected for a mean-unbiased estimator. The Pinball 5:1 model is modestly right-shifted, and the Pinball 20:1 model exhibits a pronounced positive tail - each distribution directly reflecting the quantile targeted by its respective asymmetric loss.

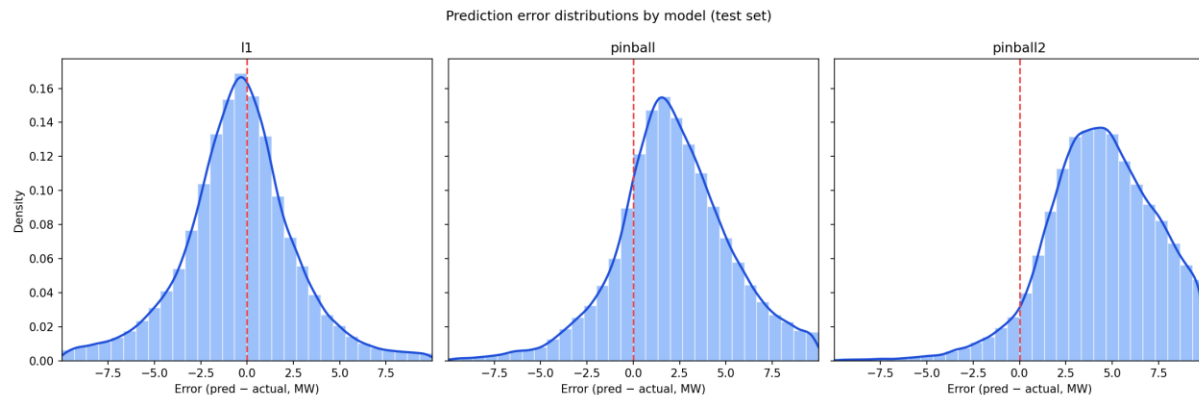


Figure 8: Distribution of test-set prediction errors (forecast minus actual, MW) for each model. The L1 model is symmetric and centred near zero. The Pinball 5:1 model shows a modest rightward shift, and the Pinball 20:1 model a pronounced positive tail, reflecting the increasingly aggressive upward bias required to minimise the cost under higher asymmetry ratios.

5. Conclusion and future work

5.1 Findings

This paper addressed day-ahead electricity demand forecasting under asymmetric operational costs using a 3-year national-scale, high-resolution dataset of electricity use prediction and actual consumption. Our main findings are:

- **Asymmetric costs are pervasive and substantial.** In electricity markets, the costs of under-predicting demand substantially exceed those of over-prediction. Empirically plausible cost ratios $c_u/c_o \in [5,20]$ have a large impact on the optimal forecasting strategy.
- **The newsvendor framework provides a principled optimality condition.** The cost-minimising procurement level corresponds to a specific quantile $\tau^* = c_u/(c_u + c_o)$ of the conditional demand distribution. For a cost ratio of 5, this yields $\tau^* \approx 0.833$, i.e. the 83.3rd percentile of demand - a systematic upward shift that is normatively *desirable* under asymmetric costs.
- **Asymmetric loss training effectively targets the cost-minimising quantile.** A feedforward neural network trained with a cost-sensitive asymmetric loss function (specifically, a non-linear variant of the pinball loss) at the appropriate quantile achieves lower expected operational cost than an L1-trained neural network baseline. The cost reduction is consistent across a wide range of synthetic cost function parameters.
- **Symmetric loss is a misleading evaluation metric under asymmetric costs.** The L1-trained neural network achieves lower MAE than the asymmetric-loss model yet incurs higher expected operational cost. Practitioners should adopt cost-sensitive metrics such as asymmetric loss or expected operational cost for model selection and evaluation.
- **Post-hoc calibration helps but does not substitute for cost-aware training.** A global affine correction fitted to the operational cost recovers much of the gap on

asymmetric metrics, but cannot match a model trained end-to-end, because it cannot adapt the upward bias to the input-conditional structure of forecast error.

- **The approach is practical and improves on the operational forecast.** Our model is a small, open-source feedforward network that trains in minutes on commodity hardware. Even though predictive accuracy was not an objective of this work, it improves on the Noga operational forecast not only on cost-sensitive metrics but also under symmetric error metrics such as MSE; we believe substantial further gains in this aspect are readily achievable. Together with a demonstrated economic benefit, this makes the method realistic to adopt in practice.

5.2 Practical implications

For market participants (electricity retailers, large consumers, aggregators), the key takeaway is that standard forecasting tools optimised for MSE are likely to systematically under-procure electricity, exposing them to avoidable real-time market risk. Incorporating asymmetric cost modelling - through direct asymmetric-loss training or post-hoc quantile adjustment - can yield meaningful reductions in procurement costs with negligible additional computational overhead. **For grid operators** (transmission system operators, independent system operators), this implies that incentivising cost-sensitive forecasting among market participants could improve overall system reliability. If all participants systematically under-forecast demand, the system operator faces persistent shortfalls in committed reserves. Regulatory frameworks that reward forecast accuracy measured by cost-sensitive metrics, rather than MSE, would align private incentives with system reliability objectives. **For policymakers**, our results support the inclusion of cost-sensitive forecasting requirements in market design rules. For example, mandatory reporting of asymmetric (pinball) loss at relevant quantiles, alongside MSE, would provide a richer picture of forecast quality that better reflects operational costs.

5.3 Limitations

Along with these positive findings, this work also has several limitations, as follows:

- **Synthetic cost functions.** Since we do not have access to actual cost data from the Israeli electricity market, all cost evaluations rely on synthetic functions that satisfy qualitative constraints. The results may not precisely reflect actual market costs.
- **Time-invariant cost asymmetry.** Although our model operates at fine (5-minute) granularity, we treat the cost asymmetry as constant across the day. In practice, markets clear at sub-hourly intervals, and cost asymmetry may vary significantly across hours of the day.
- **Point forecasts only.** Our neural network produces a single scalar forecast at the target quantile. A richer approach would estimate the full predictive distribution, enabling uncertainty quantification and adaptive quantile selection.
- **Fixed cost ratio.** We assume that c_u/c_o is constant over time. In practice, the relative costs may vary with fuel prices, demand levels, and seasonal reserve availability.

5.4 Directions for future research

- **Distributional forecasting.** Training models to output full predictive distributions (via normalising flows, conformal prediction, or Bayesian deep learning) would enable the optimal quantile to be selected adaptively as a function of current cost parameters. We regard this, rather than training against a fixed cost function, as the proper long-term formulation of the problem, for the reasons set out in the concluding remarks below.
- **Nonlinear and non-smooth cost functions.** Step-function or tiered cost structures require optimisation approaches beyond the pinball loss, such as distributionally robust optimisation or stochastic programming.
- **Online and adaptive learning.** Adaptive algorithms that update the forecasting model in real time as market conditions change would be valuable for practical deployment.

- **Richer temporal modelling and time-varying cost asymmetry.** Our model already operates at fine (5-minute) granularity but treats each target slot largely independently and assumes a constant cost ratio. Explicitly modelling intra-day and cross-day demand dynamics, for example with Transformer architectures (Vaswani et al., 2017) or temporal convolutional networks, together with a cost asymmetry that varies by hour of the day, could better exploit the structured hour-of-week error pattern documented in Section 3.1.2.
- **Multi-product forecasting.** In practice, electricity retailers must manage procurement across multiple products (day-ahead energy, reserves, ancillary services) with correlated costs and demands. A joint cost-sensitive forecasting framework would be a substantial extension.
- **Real cost data.** Collaboration with market participants or operators to obtain actual cost data would enable precise calibration and remove the need for synthetic cost functions.

5.5 Concluding remarks

Electricity demand forecasting sits at the intersection of time series analysis, optimisation under uncertainty, and energy economics. As markets become more complex, driven by variable renewable energy, demand response, and distributed storage, the costs of forecast errors are likely to increase, making cost-sensitive forecasting ever more important.

This work demonstrates that a conceptually simple modification - replacing the MSE training objective with a cost-sensitive asymmetric loss function (specifically, a non-linear variant of the pinball loss at the cost-minimising quantile) - leads to meaningful improvements in operational performance without sacrificing computational tractability.

We stress, however, that training against a *predefined, fixed* cost function is not the solution we ultimately advocate. It couples the learned model to one particular cost structure, so that a change in the cost ratio, driven by fuel prices, reserve availability, or the time of day, in principle requires retraining, and it collapses the rich uncertainty in future

demand into a single quantile chosen in advance. We view it as one useful step in the right direction rather than the destination. The desired formulation, in our view, is to *decouple estimation from decision-making*: first predict the full *conditional distribution* of demand given the available features, and then, at decision time, optimise the procurement quantity against whatever cost function currently prevails. This separation would let a single trained model serve any cost structure, adapt the optimal bias to the input-conditional uncertainty of each slot, and expose the full distribution for downstream risk-aware decisions. Realising it is the main direction of our future work. We hope these findings contribute to a broader shift towards cost-aware forecasting practices in the electricity industry and stimulate further research at the intersection of machine learning and energy systems optimisation.

6. References

- Abbasabadi, N., & Ashayeri, J. K. M. (2019). Urban energy use modeling methods and tools: A review and an outlook. *Building and Environment*, 161. <https://doi.org/10.1016/j.buildenv.2019.106270>
- Andersen, F. M., Gunkel, P. A., Jacobsen, H. K., & Kitzing, L. (2021). Residential electricity consumption and household characteristics: An econometric analysis of Danish smart-meter data. *Energy Economics*, 100. <https://doi.org/10.1016/j.eneco.2021.105341>
- Arrow, K. J., Harris, T., & Marschak, J. (1951). Optimal Inventory Policy. *Econometrica*, 19(3). <https://doi.org/10.2307/1906813>
- Ascione, F., De Masi, R. F., de Rossi, F., Fistola, R., Sasso, M., & Vanoli, G. P. (2013). Analysis and diagnosis of the energy performance of buildings and districts: Methodology, validation and development of Urban Energy Maps. *Cities*, 35, 270–283. <https://doi.org/10.1016/j.cities.2013.04.012>
- Ballarini, I., Corgnati, S. P., & Corrado, V. (2014). Use of reference buildings to assess the energy saving potentials of the residential building stock: The experience of TABULA project. *Energy Policy*, 68, 273–284. <https://doi.org/10.1016/j.enpol.2014.01.027>
- Bet-Hazavdi, E., & Drori, A. (2010). *National energy efficiency program*. Jerusalem: Ministry of National Infrastructures
- Bugin, D., Kissinger, M., & Erell, E. (2022). In Israel, too, Buildings don't use energy, people do: Analyzing lifestyles to understand household electricity consumption [in Hebrew]. *Ecology and Environment*, 13(3), 57–66.
- Chavat, J., Nesmachnow, S., Graneri, J., & Alvez, G. (2022). ECD-UY, detailed household electricity consumption dataset of Uruguay. *Sci Data*, 9(1), 21. <https://doi.org/10.1038/s41597-022-01122-x>
- Chávez, P., & Barbero, D. (2023). Methodology to characterize urban areas with similar daily electricity load curves using smart meters and census information (Montevideo-Uruguay). *Energy & Environment*. <https://doi.org/10.1177/0958305x231169011>
- Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time Series Feature Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*, 307, 72–77. <https://doi.org/10.1016/j.neucom.2018.03.067>
- Christoffersen, P. F., & Diebold, F. X. (2009). Optimal Prediction Under Asymmetric Loss. *Econometric Theory*, 13(6), 808–817. <https://doi.org/10.1017/s0266466600006277>
- de Wilde, P. (2021). Building Performance Gaps: a Commentary. *Academia Letters*. <https://doi.org/10.20935/al815>
- di Staso, U., Giovannini, L., Berti, M., Prandi, F., Cipriano, P., & De Amicis, R. (2015). *Large-Scale Residential Energy Maps: Estimation, Validation and Visualization, Project SUNSHINE - Smart Urban Services for Higher Energy Efficiency Data Management*

- Technologies and Applications: Third International Conference, DATA 2014, Vienna, Austria.
- Ferrando, M., Causone, F., Hong, T., & Chen, Y. X. (2020). Urban building energy modeling (UBEM) tools: A state-of-the-art review of bottom-up physics-based approaches. *Sustainable Cities and Society*, 62. <https://doi.org/10.1016/j.scs.2020.102408>
- Few, J., Manouseli, D., McKenna, E., Pullinger, M., Zapata-Webb, E., Elam, S., Shipworth, D., & Oreszczyn, T. (2023). The over-prediction of energy use by EPCs in Great Britain: A comparison of EPC-modelled and metered primary energy use intensity. *Energy and Buildings*, 288. <https://doi.org/10.1016/j.enbuild.2023.113024>
- Galvin, R. (2014). Making the ‘rebound effect’ more useful for performance evaluation of thermal retrofits of existing homes: Defining the ‘energy savings deficit’ and the ‘energy performance gap’. *Energy and Buildings*, 69, 515–524. <https://doi.org/10.1016/j.enbuild.2013.11.004>
- Gassar, A. A. A., & Cha, S. H. (2020). Energy prediction techniques for large-scale buildings towards a sustainable built environment: A review. *Energy and Buildings*, 224. <https://doi.org/10.1016/j.enbuild.2020.110238>
- Gunkel, P. A., Klinge Jacobsen, H., Bergaentzlé, C.-M., Scheller, F., & Møller Andersen, F. (2023). Variability in electricity consumption by category of consumer: The impact on electricity load profiles. *International Journal of Electrical Power & Energy Systems*, 147. <https://doi.org/10.1016/j.ijepes.2022.108852>
- Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3), 914–938. <https://doi.org/10.1016/j.ijforecast.2015.11.011>
- Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, 32(3), 896–913. <https://doi.org/10.1016/j.ijforecast.2016.02.001>
- Hong, T. Z., Chen, Y. X., Luo, X., Luo, N., & Lee, S. H. (2020). Ten questions on urban building energy modeling. *Building and Environment*, 168. <https://doi.org/10.1016/j.buildenv.2019.106508>
- Howard, B., Parshall, L., Thompson, J., Hammer, S., Dickinson, J., & Modi, V. (2012). Spatial distribution of urban building energy consumption by end use. *Energy and Buildings*, 45, 141–151. <https://doi.org/10.1016/j.enbuild.2011.10.061>
- Howland, A. (2013). *The Residential Energy Map: Catalyzing Energy Efficiency Through Remote Energy Assessments and Improved Data Access*. MIT.
- Johari, F., Peronato, G., Sadeghian, P., Zhao, X., & Widen, J. (2020). Urban building energy modeling: State of the art and future prospects. *Renewable & Sustainable Energy Reviews*, 128. <https://doi.org/10.1016/j.rser.2020.109902>

- Kaden, R., & Kolbe, T. H. (2013). *City-wide total energy demand estimation of buildings using semantic 3D city models and statistical data* Proceedings of ISPRS 8th Conference, Istanbul.
- Kingma, D. P., & Ba, J. (2015). *Adam: A Method for Stochastic Optimization* 3rd International Conference for Learning Representations, San Diego, CA.
- Kirschen, D., & Strbac, G. (2018). *Fundamentals of Power System Economics* (2nd ed.). John Wiley & Sons.
- Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46(1). <https://doi.org/10.2307/1913643>
- Kolter, J. Z., & Ferreira, J. (2011, 7–11.8.2011). *A Large-Scale Study on Predicting and Contextualizing Building Energy Usage* Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco.
- Lazzari, F., Mor, G., Cipriano, J., Gabaldon, E., Grillone, B., Chemisana, D., & Solsona, F. (2022). User behaviour models to forecast electricity consumption of residential customers based on smart metering data. *Energy Reports*, 8, 3680–3691. <https://doi.org/10.1016/j.egyr.2022.02.260>
- Lin, J.-L., Zhang, Y., Zhu, K., Chen, B., & Zhang, F. (2020). Asymmetric Loss Functions for Contract Capacity Optimization. *Energies*, 13(12). <https://doi.org/10.3390/en13123123>
- Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). *Accurate intelligible models with pairwise interactions* The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,
- Malik, J., Mahdavi, A., Azar, E., Putra, H. C., Berger, C., Andrews, C., & Hong, T. Z. (2022). Ten questions concerning agent-based modeling of occupant behavior for energy and environmental performance of buildings. *Building and Environment*, 217. <https://doi.org/10.1016/j.buildenv.2022.109016>
- Mathew, P. A., Dunn, L. N., Sohn, M. D., Mercado, A., Custudio, C., & Walter, T. (2015). Big-data for building energy performance: Lessons from assembling a very large national database of building energy use. *Applied Energy*, 140, 85–93. <https://doi.org/10.1016/j.apenergy.2014.11.042>
- McKenna, E., Few, J., Webborn, E., Anderson, B., Elam, S., Shipworth, D., Cooper, A., Pullinger, M., & Oreszczyn, T. (2022). Explaining daily energy demand in British housing using linked smart meter and socio-technical data in a bottom-up statistical model. *Energy and Buildings*, 258. <https://doi.org/10.1016/j.enbuild.2022.111845>
- Morales, J. M., Conejo, A. J., Madsen, H., Pinson, P., & Zugno, M. (2014). *Integrating Renewables in Electricity Markets: Operational Problems*. Springer. <https://doi.org/10.1007/978-1-4614-9411-9>

- Olivia, G.-S., & Christopher, T. A. (2015). In-use monitoring of buildings: An overview and classification of evaluation methods. *Energy and Buildings*, 86, 176–189. <https://doi.org/10.1016/j.enbuild.2014.10.005>
- Olivia, G. S., & Christopher, T. A. (2015). In-use monitoring of buildings: An overview and classification of evaluation methods. *Energy and Buildings*, 86, 176–189. <https://doi.org/10.1016/j.enbuild.2014.10.005>
- Pincetl, S., Murdock, J., Kawano, Y., & Elizabeth, Z. (2013). *Policy Factsheet: CCSC Releases First Interactive Map of Los Angeles Electricity Use*. California Center for Sustainable Communities.
- Prataviera, E., Vivian, J., Lombardo, G., & Zarrella, A. (2022). Evaluation of the impact of input uncertainty on urban building energy simulations using uncertainty and sensitivity analysis. *Applied Energy*, 311. <https://doi.org/10.1016/j.apenergy.2022.118691>
- Rothfusz, L. P. (1990). The heat index equation (or, more than you ever wanted to know about heat index). In. Fort Worth, TX Now: NWS Southern Region Headquarters.
- Shahrokni, H., Levihn, F., & Brandt, N. (2014). Big meter data analysis of the energy efficiency potential in Stockholm's building stock. *Energy and Buildings*, 78, 153–164. <https://doi.org/10.1016/j.enbuild.2014.04.017>
- Stoft, S. (2002). *Power System Economics: Designing Markets for Electricity*. IEEE Press.
- Sunikka-Blank, M., & Galvin, R. (2012). Introducing the prebound effect: the gap between performance and actual energy consumption. *Building Research & Information*, 40(3), 260—273. <https://doi.org/10.1080/09613218.2012.690952>
- Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, 72(1), 37–45. <https://doi.org/10.1080/00031305.2017.1380080>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is All You Need* 31st Conference on Neural Information Processing Systems, Long Beach, CA.
- Vecchi, F., Berardi, U., & Mutani, G. (2023). Data-driven urban building energy models for the platform of Toronto. *Energy Efficiency*, 16(4). <https://doi.org/10.1007/s12053-023-10106-8>
- Wang, Y., Gan, D., Sun, M., Zhang, N., Lu, Z., & Kang, C. (2019). Probabilistic individual load forecasting using pinball loss guided LSTM. *Applied Energy*, 235, 10–20. <https://doi.org/10.1016/j.apenergy.2018.10.078>
- Wang, Y., Jin, A. S., & Sanders, K. T. (2026). A systematic review of literature utilizing residential smart meter data. *Renewable and Sustainable Energy Reviews*, 225. <https://doi.org/10.1016/j.rser.2025.116130>
- Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4), 1030–1081. <https://doi.org/10.1016/j.ijforecast.2014.08.008>

- Westermann, P., Deb, C., Schlueter, A., & Evins, R. (2020). Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data. *Applied Energy*, 264. <https://doi.org/10.1016/j.apenergy.2020.114715>
- Wu, J., Wang, Y.-G., Tian, Y.-C., Burrage, K., & Cao, T. (2021). Support vector regression with asymmetric loss for optimal electric load forecasting. *Energy*, 223. <https://doi.org/10.1016/j.energy.2021.119969>
- Zhang, J., Wang, Y., & Hug, G. (2022). Cost-oriented load forecasting. *Electric Power Systems Research*, 205. <https://doi.org/10.1016/j.epsr.2021.107723>