UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Elizaveta Nikolaeva

Analyzing the solar energy potential of Smart Cities

Master's Thesis (30 ECTS)

Supervisor: Pelle Jakovits, PhD

Tartu 2023

Analyzing the solar energy potential of Smart Cities

Abstract:
Smart cities, driven by technology and innovation, aim to enhance the quality of life for citizens while prioritizing sustainable urban development. Embracing renewable energy sources such as solar energy is a vital strategy contributing to the sustainability of a city. For this reason, understanding the solar energy production potential and identifying consumption to production balance becomes crucial for achieving sustainable urban development goals. This thesis aims to employ a data-driven approach for analyzing the estimated solar energy production from buildings' roofs in Tartu city, validate and refine previous estimations, and explore methods to predict energy consumption in buildings. The purpose of the analysis and investigations is to provide meaningful insights on Tartu's solar potential and facilitate future assessments of energy balance within the city. The revised approach to assessing energy production demonstrated better accuracy, while the implementation of data cleaning and cache utilization resulted in a 5-fold increase in processing speed. Furthermore, there were provided valuable recommendations regarding optimal roof characteristics specific to the region for enhanced solar panels productivity. In addition, the thesis suggests which models can be suitable for predicting energy consumption in urban buildings and defines requirements on what data should be collected in the future to be able to realise such predictions in Tartu.

Keywords:
Data-driven analysis, Solar Potential Estimation, Energy balance, City Sustainability, Smart City

CERCS:
P170 Computer science, numerical analysis, systems, control

Tarkade linnade päikeseenergia potentsiaali analüüsimine

Lühikokkuvõte:

Tehnoloogiast ja innovatsioonist juhitud nutikate linnade eesmärk on tõsta kodanike elukvaliteeti, seades esikohale säästva linnaarengu. Taastuvate energiaallikate, nagu päikeseenergia, omaksvõtmine on linna jätkusuutlikkuse tagamisel ülitähtis strateegia. Sel põhjusel muutub päikeseenergia tootmispotentsiaali mõistmine ja tarbimise ja tootmise tasakaalu kindlakstegemine säästva linnaarengu eesmärkide saavutamiseks ülioluliseks. Käesoleva lõputöö eesmärk on kasutada andmepõhist lähenemist, et analüüsida hinnangulist päikeseenergia tootmist Tartu linna hoonete katustelt, kinnitada ja täpsustada varasemaid hinnanguid ning uurida meetodeid energiatarbimise prognoosimiseks hoonetes. Analüüsi ja uuringute eesmärk on anda sisukaid teadmisi Tartu päikesepotentsiaalist ning hõlbustada linna energiabilansi edaspidist hindamist. Energiatootmise hindamise muudetud lähenemisviis näitas paremat täpsust, samas kui andmete puhastamise ja vahemälu kasutamise tulemusel suurenes töötlemiskiirus 5 korda. Lisaks anti väärtuslikke soovitusi piirkonnale omaste optimaalsete katuseomaduste kohta päikesepaneelide tootlikkuse suurendamiseks. Lisaks pakutakse lõputöös välja, millised mudelid võivad sobida linnahoonete energiatarbimise prognoosimiseks ning määratletakse nõuded, milliseid andmeid tuleks tulevikus koguda, et selliseid ennustusi Tartus realiseerida.

Võtmesõnad:

Andmepõhine analüüs, päikesepotentsiaali hindamine, energiabilanss, linna jätkusuutlikkus, tark linn

CERCS:

P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Contents

List of Figures

List of Tables

1 Introduction

In pursuit of sustainable urban development, smart cities have become the answer to the ever-increasing challenges of rapid urbanisation and environmental concerns. These cities utilise technology, data, and innovative approaches to enhance the quality of life for citizens while minimizing their ecological footprint. There are different ways of achieving these goals from switching to ecological transport modes to reducing and managing waste. Transitioning to renewable energy (RE) is crucial in achieving sustainability, and solar energy emerges as a particularly promising form of RE, playing a pivotal role in this endeavor [19].

With the increasing adoption of solar energy in urban environments, there arises a growing need to assess the potential energy production when implementing rooftop photovoltaic (PV) systems and evaluate the coverage of energy consumption they provide. Determining the optimal integration of these systems within a city requires the estimation of the energy output achievable with their installation. However, simply assessing the solar potential is not sufficient for a comprehensive analysis, as this does not identify the main criteria for maximising the efficiency of rooftop PV systems. Conducting such an analysis is a necessary step to establish an accurate energy balance for the city. This valuable examination not only sheds light on sustainability trends but also offers useful insights for future construction. At present, sustainability has become a pressing concern, which is evident from the EU's commitment to developing greener and more environmentally conscious cities [7, 8, 9]. By understanding the city's energy needs and potential energy production through solar sources, we can make informed decisions to promote sustainable growth and development.

The primary objective of this thesis is to utilize a data-driven approach to analyze the potential solar energy generation from rooftops in Tartu city, while validating and enhancing existing estimations and exploring suitable approaches for predicting buildings' energy use. It is essential to note that the existing solar productivity estimation is derived from Bohdan Romashchenko's master's thesis [20] the content of which is described in the Section 2 of this manuscript. In Romashchenko's work, the accuracy of these estimations could not be assessed due to the absence of data on real PV installations in Tartu. Therefore, validation becomes crucial to ensure the reliability of the obtained production values and the correctness of the analysis based on this data. Moreover, understanding the factors that drive solar potential variations among different buildings is vital for practical applications. For instance, buildings constructed during different time periods or in different districts may possess varying roof features directly affecting energy production. Conducting a comprehensive analysis and investigation of these factors will provide valuable insights into Tartu's solar energy capabilities and inform strategies for future improvements. Besides, in order to guide urban planners to design energy efficient buildings and improve the overall sustainability of the city, an understanding of the balance of energy production and consumption in the city is needed. However, while less granular consumption (city level or city area level) might be possible to acquire, access to energy consumption data of buildings, especially at the city scale, is difficult because it is considered confidential. As a result, it is essential to explore approaches for predicting energy consumption, especially based on partially available information, and identify the necessary data for such estimations, laying the groundwork for future energy balance assessments within the city. Based on these challenges and objectives, the following research questions are formulated:

1. Solar energy production potential
    (a) **RQ1:** How accurate is the solar energy potential estimated in the previous work comparing it to the actual deployments in the city?
    (b) **RQ2:** What types of buildings (based on their roofs' features) contribute more to the energy potential? Which types of constructions should be prioritised in the future to improve energy potential of the city?
    (c) **RQ3:** Which city areas contribute most to the energy potential?
2. Energy consumption estimation
    (a) **RQ1:** What are the suitable models for energy consumption estimation?
    (b) **RQ2:** Is it possible to estimate the monthly energy consumption of city buildings based on partial information, which is currently available or can be gathered in short time?
    (c) **RQ3:** Which data should be collected to perform such estimations?

To answer the research questions, it was first important to gather the data on actual solar energy production from rooftop PV installations in Tartu. Afterwards, previous estimations were validated by comparing estimated and real production values per square meter of installed panels. Since the comparison revealed substantial differences, a thorough investigation was conducted to define the underlying factors contributing to these variations. In the subsequent steps, refinement of the initial estimations as well as data cleaning were performed followed by the validation and analysis of new results. Data analysis step answers research questions RQ1.1 to 1.3. Finally, a careful examination of the domain literature along with data searching and exploration were conducted to address the research questions RQ2.1 to RQ2.3.

The reminder of the thesis is structured in a following way. Chapter 2 provides a background information on the analysis of potential energy consumption while giving a short overview of the Bohdan Romashchenko's master's thesis. Moreover, it explains challenges related to energy consumption assessment and describes of the related works in this domain. In chapter 3, methodology of this thesis is explained including the description of datasets, conducted analysis and refinement of the previous work, and problems encountered. Chapter 4 contains a more detailed information on the energy consumption prediction approaches and evaluates Tartu's energy balance based on available data. Chapter 5 covers the main results. In chapter 6, the work is discussed, research questions are answered and directions for future work are provided, whereas chapter 7 summarizes the work done and concludes the thesis.

2 Background

This chapter delves into the significance of analyzing the potential for solar energy production and estimating energy consumption in urban areas. Firstly, it presents a review of the previous research upon which this thesis builds, followed by discussions on potential improvements to the existing approaches. Additionally, the chapter explores the challenges of accessing energy consumption data due to privacy concerns and provides a concise overview of the prediction models used for energy consumption estimation.

2.1 Analysis of solar energy production potential

Accurate estimation of solar energy production potential is crucial for promoting sustainable development and advancing the adoption of renewable energy sources in cities. Having such data and extracting useful knowledge from it can help policymakers and urban planners to make well-informed decisions regarding the integration of solar power in urban areas. It allows for the identification of locations where PV system installations are most feasible and can yield maximum benefits. Moreover, it can become clearer how to design new buildings or retrofit existing structures to be more energy-efficient, contributing to the overall sustainability of the urban environment. With this information, policymakers can design effective solar energy policies, such as feed-in tariffs, tax credits, or grants, to encourage investment in solar power projects. Overall, aligning policies, infrastructure development, and investment decisions with analyzed solar potential data drives the transition towards cleaner and more sustainable energy sources, reducing carbon emissions and contributing to a greener urban future.

Solar production estimation of rooftop PV systems is based on a combination of several factors including location, tilt, orientation, peak power, weather, shading, etc. According to the paper by S. Freitas *et al.* [12], the approaches for calculation of the potential can be classified as empirical and computational. The former use data from meteorological stations to estimate solar irradiance. They typically measure global and diffuse irradiation on the horizontal plane and then use formulas to calculate the direct component based on the sun's zenith angle. However, to estimate the irradiance on tilted surfaces, additional geometric formulations are required, which can consider cloud cover and other factors affecting solar radiation. These models provide valuable estimates for different surfaces but may require assumptions about diffuse and ground-reflected radiation. Computational solar radiation models are advanced numerical simulations that calculate solar irradiance and radiation in complex and dynamic environments, such as urban areas with obstructions and varying topographies. There are various solutions and simulation software implementing computational approach such as Daysim[1], ArcGIS

[1] Daylighting analysis software: `https://daysim.software.informer.com/4.0/`. [10.08.2023]

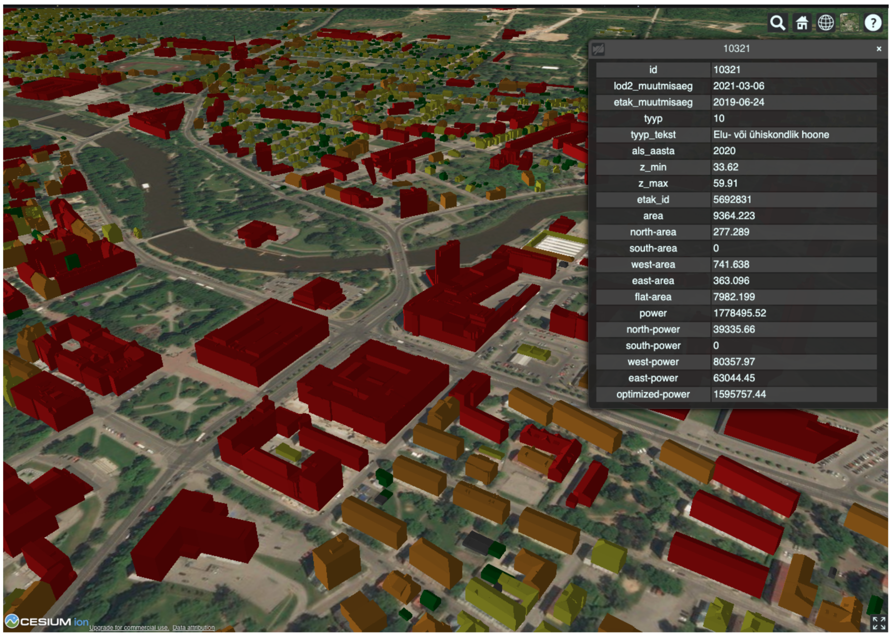
Solar Analyst[2], or r.sun[3]. Unlike empirical models, computational models employ sophisticated algorithms and mathematical equations to account for the interactions of sunlight with buildings, terrain, and atmospheric conditions. These models provide a more detailed and accurate representation of solar potential by considering factors such as shading or reflections [12].

As was mentioned in the chapter 1, this research is related to Bohdan Romashchenko's master's thesis [20], where he aimed to create software that would estimate solar potential of any Estonian city and visualize the results, using the data provided by Estonian Land Board. The potential energy production was assessed via PVGIS application programming interface (API) which is part of a web tool created by EU Science Hub for providing information on solar irradiation and PV system performance [3]. PVGIS refers to computational approaches. It calculates solar radiation by using satellite-based data and models that take into account the sun's position, weather, and shading effects from terrain. PVGIS tool estimates PV power output by combining solar radiation data with information about the PV system's characteristics, such as panel efficiency, tilt angle, orientation, etc. As a result of his thesis, Bohdan successfully created a solution for extracting geographical features and estimating solar power production in Tartu. Additionally, he developed a user-friendly web application (see Figure 1) to visualize and present the results effectively. However, while the Bohdan's work provided a well-designed mapping software and codebase for communicating with the API, it also had certain limitations and room for future improvements.

One significant challenge was the lack of data for the validation of whether the author's estimations were close to reality. Validation is an essential step in such assessment since it ensures the accuracy and reliability of the results. Although estimations can sometimes deviate from real production values due to unpredictable weather conditions, the discrepancies per measurement unit should be minimized to have a proper understanding of solar performance. Another significant issue was the prolonged processing time (65 minutes) for API requests, which made it more difficult in cases where the code needed to be run again. Addressing these limitations by validating the estimations with real solar production data and implementing improvements to the request processing would not only enhance our understanding of the city's potential for sustainable development but also improve the software's usability. Lower computation time is especially crucial for applying the methodology to much larger cities than Tartu. In addition, whereas having a visualization of solar production potential is undeniably valuable, conducting data analysis would provide meaningful insights for achieving sustainable urban development goals. By delving into the solar production data, we can identify specific buildings or city areas that could have more contribution to the overall solar energy output. Furthermore,

[2] Solar radiation analysis software: `https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/area-solar-radiation.htm`. [10.08.2023]

[3] Solar irradiance and irradiation model: `https://grass.osgeo.org/grass82/manuals/r.sun.html`. [10.08.2023]



10321

id	10321
lod2_muutmisaeg	2021-03-06
etak_muutmisaeg	2019-06-24
tyyp	10
tyyp_tekst	Elu- või ühiskondlik hoone
als_aasta	2020
z_min	33.62
z_max	59.91
etak_id	5692831
area	9364.223
north-area	277.289
south-area	0
west-area	741.638
east-area	363.096
flat-area	7982.199
power	1778495.52
north-power	39335.66
south-power	0
west-power	80357.97
east-power	63044.45
optimized-power	1595757.44

Figure 1. Example of web application interface visualizing Tartu's solar power production [20].

this analysis can reveal the characteristics of rooftops that are better suited for solar panel installations. With the careful examination of factors such as roof orientation or tilt, we can determine which rooftops are optimal for PV systems. For instance, south-facing rooftops with a certain tilt angle could capture more sunlight throughout the day, maximizing energy generation potential. In addition, it could highlight the use of sloped systems on flat roofs to increase solar energy production capabilities.

The problems that arose from Bogdan Romashchenko's work further shaped the current approach. The need for the validation of estimations became apparent, leading to the necessity of finding real solar energy production data. In addition, long request processing times revealed the need to improve computational efficiency, which motivated the use of optimisation and data cleaning techniques in this thesis. Furthermore, understanding the limitations of potential solar production visualisation alone, data analysis was introduced. This decision is driven by the realisation that a deeper examination of solar energy production data enables the identification of optimal roofs and helps to achieve sustainable development goals for the city. The insights gained from Bogdan Romashchenko's thesis laid the foundation for the comprehensive approach in this study. By addressing issues of validation, computational efficiency, and data analysis, this research improves the assessment of solar energy potential and contributes to the creation of a more sustainable urban future.

2.2 Energy consumption estimation

Assessing the potential energy production of rooftop photovoltaic systems is clearly significant, but estimating and understanding energy consumption is equally important for establishing an accurate energy balance and making data-driven choices for achieving sustainable urban development. However, analysis of consumption presents a number of challenges related to data availability and confidentiality. While solar energy production can be estimated based on geographical and meteorological data, energy consumption data is very limited and difficult to access. This limitation is primarily due to privacy concerns, as individual energy consumption data is sensitive and subject to various regulations protecting the collection and sharing of such data. Data relating to specific individuals or households may reveal details about a person's daily habits, their lifestyle and perhaps even their presence or absence from their home, which may have privacy implications. At the same time, businesses may need to keep their energy consumption data confidential because they could be bound by certain agreements or concerned about disclosure of information on their operational and production processes, as well as fluctuations in production volumes or occupation. As a result, energy consumption data cannot be accessed at a city-scale, which creates the need to predict the energy usage.

On a broader level, energy consumption prediction models can be classified into engineering and statistical or data-driven approaches. Engineering methods are based on physical principles used for the calculation of the energy consumption of entire buildings or their components. These methods employ sophisticated calculations and require extensive building and environmental parameters, which makes them accurate but complex and resource-intensive [24]. On the other hand, data-driven models, such as regression and artificial neural networks (ANNs), can be applied more easily and widely since they are more effective in terms of resources and require less detailed data. With this approach, historical consumption data and, depending on the model, various additional features such as buildings characteristics or weather are used to determine independent variables influencing energy consumption. This happens during so-called training process which relies on minimizing the error between actual and predicted values of dependent variable [23]. However, the lack of real energy usage data makes it difficult to estimate consumption for all buildings in the city and assess how well consumption can be covered by the production of solar energy. This thesis investigates data-driven models to find out if their application is feasible for Tartu city, and formulates research questions RQ2.1 to 2.3 in the chapter 1 in response to the challenges associated with these models.

In the paper by Nelson Fumo and M.A. Rafe Biswas [13], the authors highlight that the linear regression analysis is a promising technique for energy consumption prediction because it provides reasonable accuracy while being a relatively simple model in comparison with other methods such as ANNs or genetic algorithms. The overall goal of the paper is to provide a comprehensive reference for future studies on regression analyses for predicting energy consumption in residential buildings. In the literature review, Fumo and Biswas cover different regression models used for this purpose, including studies that examined heating energy consumption, energy performance analysis, impact of solar radiation on natural gas consumption, evaluation of heat pumps for space heating, modeling residential energy consumption at the zip-code level, and more. Each study utilized regression analysis with specific predictor variables to estimate energy consumption in different contexts, ranging from specific building characteristics to socio-economic and behavioral factors. Regarding the practical contribution, the authors implemented various regression models on hourly and daily data from a research house and found that the time interval of the observed data significantly impacted model quality. Daily interval models always performed better than hourly ones due to the fact that lower time resolution, which results from averaging over longer time periods, reduces the impact of discrepancies among individual effects [13], leading to improved model accuracy.

The research by Tao Hong *et. al* [15] discusses the modeling and forecasting of short-term electric load by exploring multiple linear regression. The proposed approach investigates the causal factors affecting electric load fluctuations at different time periods. The study analyzes yearly, monthly, daily, and hourly data, considering temperature, human activities, and seasonality as potential influencing factors. From the analysis, the paper finds an overall increasing trend in electricity consumption, potentially influenced by temperature and human activities. Seasonal patterns are evident with peak loads occurring in both winter and summer, indicating heating and cooling demands. By considering various interactions, the proposed model offers insights into seasonal variations. It also demonstrates that electricity is primarily used for cooling during summer and for heating during winter. Moreover, the study reveals that temperature and human activities significantly affect load variations at different time scales, including monthly and hourly resolutions.

One of the articles describing an interesting approach for energy consumption prediction was the research by Alessio Mastrucci *et. al* [18], aimed to identify the real energy consumption pattern and potential energy savings in a large housing stock. The focus has been on building energy retrofits as a strategy to reduce urban carbon emissions. To assess the impact of these measures, the authors developed a statistical model based on multiple linear regression incorporating various factors such as housing classification, construction year, floor area and number of occupants. By downscaling the aggregated data at the postcode level to the level of individual dwellings, energy consumption was estimated and allocated to different end-users, taking into account weather variations. With such technique, training data does not need to have building-level granularity for the accurate energy consumption prediction, thus taking into account privacy issues. The results of the study provided valuable insights to support sustainable urban design and prioritise energy modernisation measures in Rotterdam. Moreover, the methodology's generic nature suggests that it can be adapted for application in other cities and contexts.

Being able to find a substantial amount of real energy consumption data, Constantine E. Kontokosta and Christopher Tull [17] used a different approach in their modelling. The primary goal of the research was to create a forecasting model for energy consumption at various levels, including individual buildings, districts, and entire cities. This was achieved by utilizing data from energy disclosure policies and property information. The authors managed to predict electricity and natural gas usage for 1.1 million buildings in New York City and assessed the accuracy of the models using real consumption data. In pursuit of this objective, the researchers employed various statistical models, including linear regression (OLS), random forest, and support vector regression (SVM). The study's significant findings revealed that the OLS model proved most effective in predicting energy use for the entire city, while the SVM model exhibited the lowest mean absolute error when forecasting energy usage specifically within the subset of buildings required to report data. Additionally, the study sought to identify significant predictors of energy use, such as building age, type, size, and shared party walls. Overall, the research provided valuable insights on urban energy analytics and energy policy, facilitating informed decision-making for efficient energy usage and sustainability goals.

Apart from regression models, there are other more sophisticated methods for energy consumption modelling and forecasting. Artificial neural networks, as well as their subclass deep neural networks (DNNs), are widely used for this matter. They represent nonlinear machine learning models that mimic the behaviour of human bran in a simplified manner to propagate information and find correlations between data points [6]. ANNs and DNNs are usually used to identify complex patterns in time series. For example, in the article by Jian Qi Wang *et. al* [22], the authors presented a novel approach using Long Short-Term Memory network to forecast periodic energy consumption. Some of the researchers suggest to use different techniques withing deep learning models together. Tae-Young Kim and Sung-Bae Cho, for instance, introduced a combination of convolutional neural networks and long short-term memory approaches in their work [16]. With this approach, they were able to achieve stable prediction of the amount of electricity consumption in a real house, while analysing the variables of household appliances that affect the prediction of energy consumption. Although ANNs and DNNs can provide high accuracy and be used for comprehensive analysis, they require a substantial amount of data for training, which can present a challenge in the context of energy consumption prediction.

In the reviewed literature, the promising nature of regression models for energy consumption prediction was highlighted as they present a relatively simple technique and provide reasonable accuracy in this research area. The most influential paper by Alessio Mastrucci *et. al* provided a methodology for downscaling energy consumption while preserving data privacy and not requiring granular building-level data for city-scale generalization. By using aggregated data at the postcode level and considering various factors like housing classification, construction year, and floor area, they achieved accurate energy consumption estimates at an individual dwelling level. This approach proved valuable for supporting sustainable urban design and energy modernization strategies in Rotterdam, offering potential applications in other cities. Constantine E. Kontokosta and Christopher Tull presented an alternative approach for the cases when sufficient amount of building-level data is available and demonstrated the feasibility of predicting energy consumption for buildings of different types. Furthermore, Tao Hong *et. al* highlighted the importance of including outdoor temperature in the model for monthly predictions. Overall, these works have influenced the author's suggestions for suitable models and data outlined in the subsection 6.1.

3 Methodology

This chapter provides a description of the main stages of this thesis, which are crucial for achieving the set goals and addressing the research questions. The chapter begins with the validation of previous estimations of Tartu's solar energy potential. Additionally, it includes an overview of encountered problems along with their resolution. All the data used during the research are also described in this chapter.

3.1 Validation of estimated solar potential data

Due to the lack of data on the output of installed PV systems in Tartu, the estimations for solar potential could not be properly validated in the previous work by Bohdan Romashchenko. During the work on this thesis, it became possible to acquire such data for the buildings renovated under the SmartEnCity project. The project received funding from the European Union's Horizon 2020 research and innovation program and had the objective of implementing a comprehensive strategy for transforming European cities into sustainable, intelligent and resource-efficient urban spaces [5]. A total of 18 buildings built in Tartu in the 1950s and 1960s became energy-efficient with the various retrofitting measures and smart solutions such as insulation, triple-glazed windows, smart meters, rooftop PV systems, etc.

3.1.1 Real solar energy production data

The actual solar energy production data was obtained from the Cumulocity platform, facilitated by the thesis' supervisor, who gained access to the platform for this purpose. 612 *csv* files were downloaded and represented the monthly cumulative solar panels output data of 17 buildings for three years, namely 2020, 2021, and 2022. Figure 2 demonstrates the percentages of empty files for each year. In 2020, there was almost no data recorded, while 2021 and 2022 had around 22% and 17% of missing data respectively. Year 2020 only had data for 3 months, namely October, November, and December. In addition, for one of the buildings, there was no data at all. Thus, after removing empty values from the dataset, the data for 16 buildings was left for further analysis.

Before performing any calculations, it needs to be checked whether the data has to be cleaned. The distribution plot of the cumulative energy production expressed in kWh is depicted in the Figure 3 and shows that the data follows a normal distribution, but there are some outliers. To filter out the outliers, lower and upper limits, which represent three standard deviations below and above the mean, were calculated:

- Lower limit: 352.70 kWh;
- Upper limit: 130249.13 kWh.

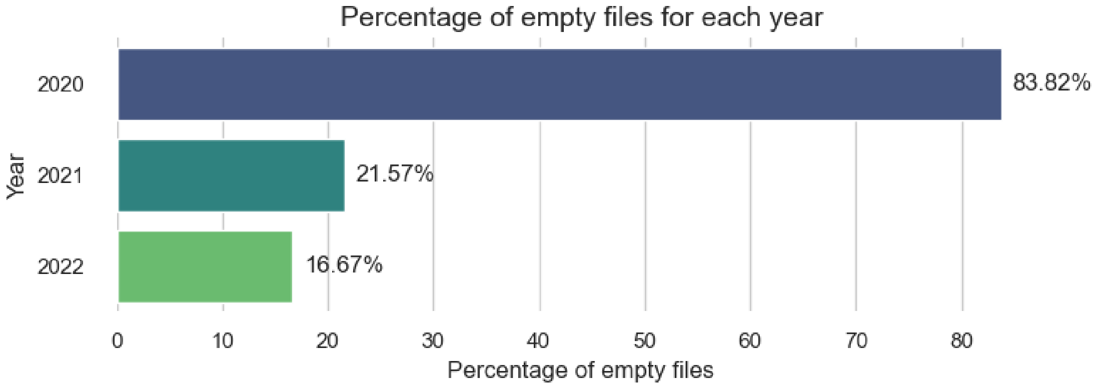


Figure 2. Percentage of empty solar panels output files for each of three years.

It was assumed that the values which are close to around 140000 kWh might be real in our case since at least some of the buildings had flat roofs, which could potentially have optimally installed solar panels. To check it, it was decided to find out which building or buildings have production values higher than upper limit. One of the buildings satisfied this condition having high last measured values per each year (see Table 1). It can be seen that towards the end or right at the end of each year, this building had fairly high cumulative production numbers. Due to this reason, the values higher than upper limit were not dropped from the dataset. On the other hand, the values lower than lower limit appeared to be very small and thus were filtered out as outliers. In total, approximately 24% of all non-empty files contained outliers.

Based on the cumulative data from non-empty files, the calculation of monthly solar production values was divided into two steps:

1. The first measurements of the current and next months are subtracted if these months are consecutive. If the months are not consecutive, the production value is set to NaN.

2. NaN values obtained from the first step are replaced with the production values calculated by subtracting last and first measurements of the corresponding file.

Table 1. The production of building having high last measurements per each year.

year	production, kWh	month
2020	62010.70	12
2021	98319.60	9
2022	145369.81	12

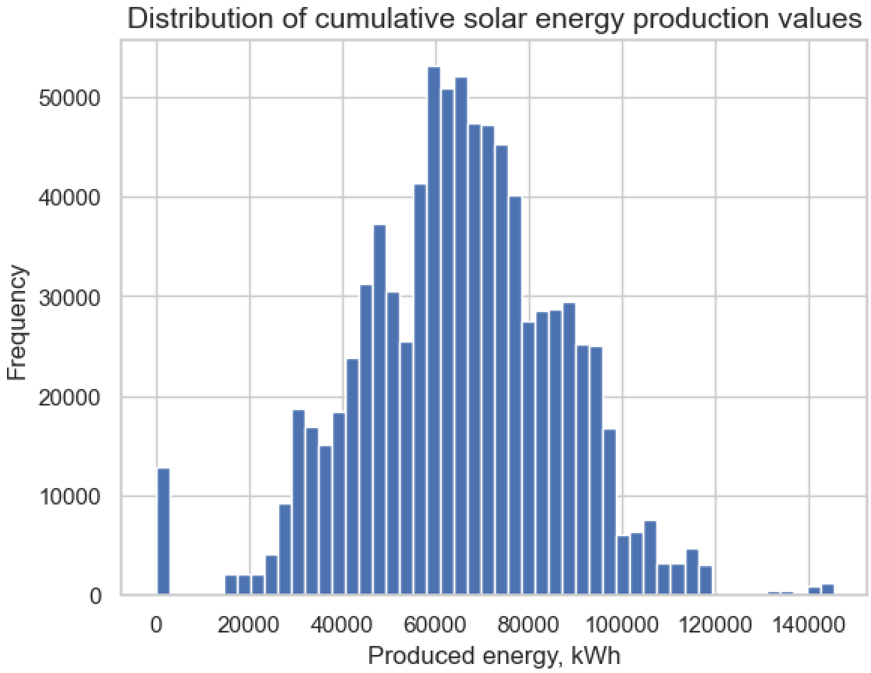


Figure 3. The distribution of the real cumulative energy production values.

After exploring some files manually, it was clear that some of them do not contain the full month data. For example, files starting on January 10 and ending on January 20. Files like these were not used for monthly energy production calculation. Data points where the day of first measurement was higher than 5 and the day of last measurement was less than 25 were dropped. These numbers were defined rather intuitively, but also by looking at the frequency of days of the months for both first and last measurements of files (see Figure 4). These thresholds are not very strict and allowed for the calculation of approximate monthly solar energy production from cumulative values. Calculation of the monthly production values showed that no buildings have solar production data available for all the 12 months, and for 10 out of 16 buildings, some years had the production calculated for at least for 10 months. Annual production was calculated separately, using an approach similar to that for monthly production. Calculation was performed by subtracting either the first measurements of consecutive years or the first and last measurements of the year. Mean energy production for each month is demonstrated in the Figure 5 and was plotted to see which months should not be neglected in terms of contribution to energy production. The graph shows that on average the production figures in January and February are more than twice as low as in March. This means that to calculate annual production, the first measurement should come from a file representing data recorded no later than March. Regarding the last measurement, it can be obtained in any month starting in October, since after that month there is no big change in energy production. With this approach, it was possible to calculate actual annual energy production for 10 and 13 buildings in 2021 and 2022 respectively.

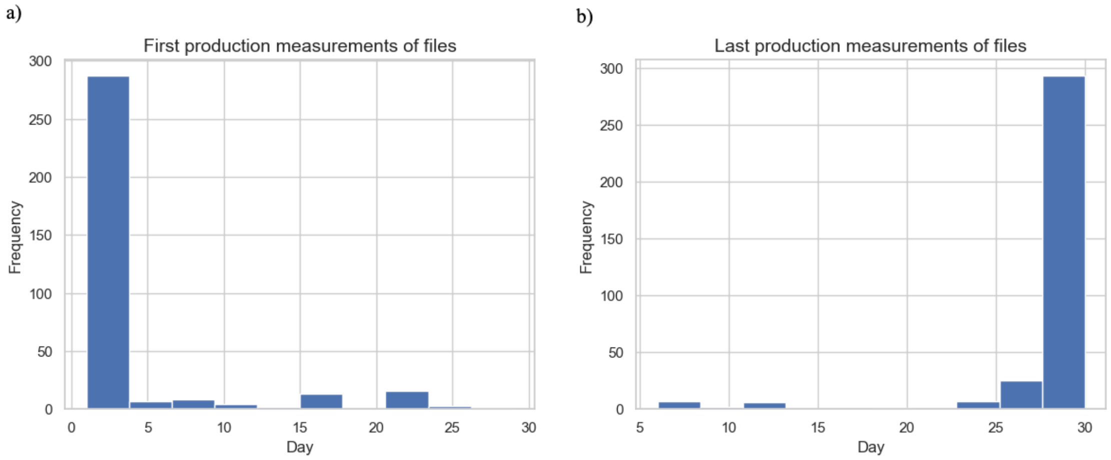


Figure 4. a) Days frequency of the first production measurements in files; b) Days frequency of the last production measurements in files.

3.1.2 Estimated solar energy potential data

Data on the estimated solar energy potential was obtained from Bohdan Romashchenko's work output[4] in *json* format. Each entry in the dataset is identified by a unique building ID and contains details about multiple roof parts that belong to the same building rooftop. This data represents information about different roofs and their associated properties:

- roof area,
- orientation,
- geographical location (longitude and latitude),
- list of points representing the shape of a particular part of the roof in EPSG 3301 coordinate system,
- tilt,
- azimuth,

[4] Previously estimated solar energy potential dataset: https://github.com/boroma4/solar-potential-output-example/blob/main/tartu-output/city-attributes.json

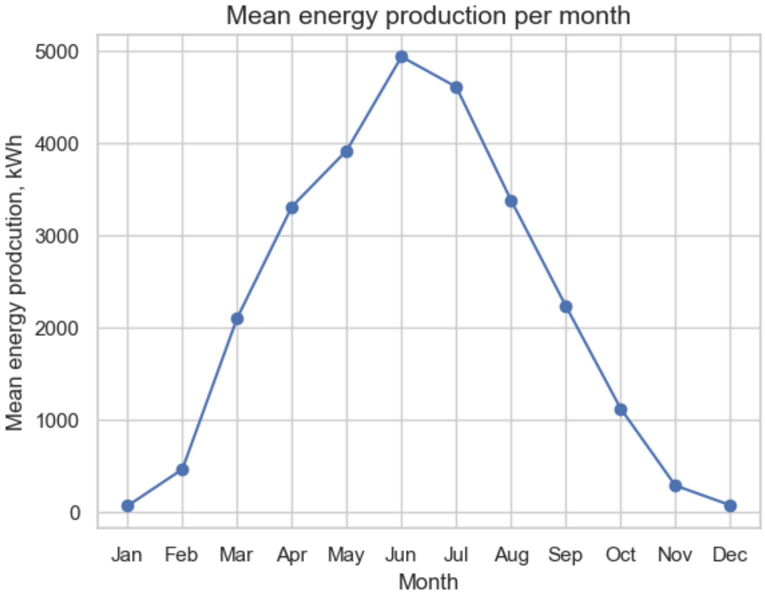


Figure 5. Mean actual solar energy production per month.

- annual solar energy production in kWh,
- monthly average solar energy production in kWh,
- list of energy production values in kWh for each month of the year,
- total energy loss in percentage.

Since the file contained nested structures, the author decided to flatten it to a Pandas dataframe which is a convenient tabular data structure of Python Pandas library [2] for managing and analysing data. There are around 79443 roof parts in the dataset which were grouped by building ID and aggregated to 20729 constructions. Each building then was mapped with the district it is located in. To perform the mapping, coordinates of the borders of each district were requested from a Tartu's ArcGIS REST Service[5], and then each building was assigned with the corresponding district by checking whether its coordinates lie inside of the polygon in question. If it was not possible to define such a polygon for a particular building, than the building was mapped with the closest district.

[5] Districs information REST API: `https://gis.tartulv.ee/arcgis/rest/services/Planeeringud/GI_linnaosad/MapServer`. [23.07.2024]

3.1.3 Comparing actual and estimated solar energy potential

To perform the validation of the estimated solar energy potential, we need to compare monthly and annual production values in kWh per square meter of installed panels with real solar potential data. For the real installations, the author calculated the area of the roof allocated for PV system manually by using an Estonian Land Information web map[6] constructed of processed aerial images or orthophotos and measurement tools available in the application. Figure 6 demonstrates an example of such photo on the map, where the panels are clearly seen. After measuring the PV systems area for each of the 16 houses, the actual production values were divided by the corresponding area. Regarding the estimations of solar potential, in the previous work, they were performed assuming that 90% of a roof area can be covered by solar panels and the dataset contained roof area values, thus monthly and annual production in kWh/m^2 could be easily calculated.

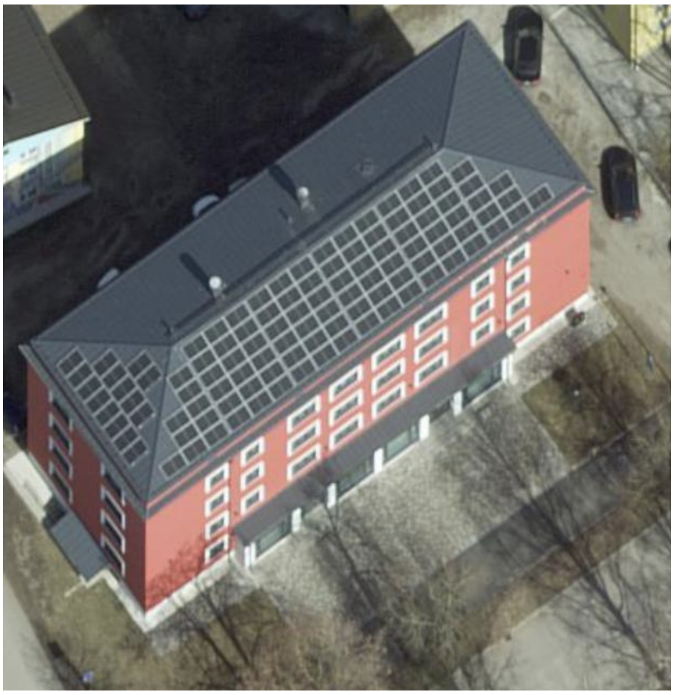


Figure 6. Orthophoto of a building at Tiigi 19, Tartu, from the Estonian Land Information web map.

As estimations for solar energy production were performed in 2022, they were only compared to the actual production values for that year. This comparison is illustrated by Table 2, where *prod* columns represent production and suffixes *a* and *e* stand for

[6] Estonian Land Information Web Map: https://xgis.maaamet.ee/xgis2/page/app/maainfo. [23.07.2023]

actual and estimated respectively. Mean Absolute Percentage Error (MAPE) for annual production was 23.61%. This error is quite big which could be due to the fact that solar panels efficiency and loss used as some of the input values for the PVGIS API did not match the values for the PV systems installed on the actual rooftops. Monthly production was also compared, but since there were multiple empty values in December, it was not accounted for during the calculations. From table 3, which represents MAPE by season, it can be seen spring months account for the best accuracy, followed by summer months, while winter months have an error of almost 120%. This might have happened if, for example, the actual weather conditions differ from what was expected by the API. Solar panels could be also turned off for a certain period knowing that, due to the cold temperatures and small number of daylight hours in the area, there will be very little energy produced. It should also be taken into account that the amount of data for validation is quite small to be able to have an accurate understanding of accuracy for the whole dataset, especially for winter months with a lot of missing values. Overall, there are discrepancies in the actual and estimated results which may stem from various factors, including different characteristics of PV systems or errors in calculations.

Table 2. Comparison of actual and estimated annual solar energy production per m^2 for 2022.

address	pv_area_a, m^2	pv_area_e, m^2	prod_a, kWh/m^2	prod_e, kWh/m^2
Tiigi 19	195.82	453.978	169.26	187.66
Tähe 2	182.69	461.862	166.30	199.47
Aleksandri 3	146.65	468.828	158.04	175.85
Turu 9	147.85	482.049	160.65	176.14
Turu 3	165.44	477.837	149.79	211.86
Pepleri 3	139.91	428.364	142.09	176.30
Pepleri 12	171.77	727.407	155.80	176.54
J.Kuperjanovi 2	193.61	583.533	149.63	178.12
Lutsu 16	187.23	427.428	107.96	177.57
Tiigi 21	161.09	423.891	144.85	175.03
Tiigi 23	162.00	433.629	159.07	175.25
Aleksandri 12	260.11	718.956	175.62	175.90
Kalevi 8	168.47	555.345	109.66	177.31

3.2 Exploring rooftops of Tartu's SmartEnCity buildings

In addition to assessing the accuracy of the estimated solar potential, the validation process revealed that actual solar panels cover significantly less than the previously assumed 90% of the roof area. This earlier value was chosen to represent the maximum

Table 3. Mean Absolute Percentage Error for each season.

Season	MAPE, %
Winter	119.08
Spring	15.27
Summer	19.16
Autumn	31.93

solar production potential, but it is overly optimistic for real-life scenarios. To achieve more realistic solar production estimates, it was decided to improve the selection of roof parts. Unlike the previous approach, which included all parts of each rooftop, this thesis suggests applying filtering to refine the selection process. By examining the sections of rooftops with existing solar panels, valuable insights can be gained to create the effective refinement strategy. Based on available data, it can be defined how the roof sections that were chosen for the installations are oriented. It is also important to identify the actual size of surface allocated to existing PV installations to better understand which percentage of roof area is available on average.

To determine the orientation of roof parts accurately, the roofs of the 17 buildings described in the Subsection 3.1.1 were plotted and compared to their orthophotos. During this process, a mistake in the calculation of roof sections' azimuths was discovered due to the incorrect axes orientation. The visualisation used in previous work did not colour the roof sections separately but coloured the whole roof, making the issue difficult to detect at the initial stage. Consequently, the rooftops' azimuths were recalculated and mapped with new orientations initially focusing on the 17 SmartEnCity buildings. Later, the recalculations were extended to cover the entire dataset[7]. Figure 7 illustrates examples of the comparisons of old and new orientations in the following manner:

- row *a* shows graphs of rooftops and the orientations assigned to their parts in the previous work,
- row *b* shows corresponding orthophotos,
- row *c* shows graphs of rooftops and the orientations assigned to their parts in the current thesis.

From the figure, it is clear that the orientations mapped in the previous work do no match the actual orientations although the coordinates and graphs axes placement were proved to be correct based on the use of the web map in this work. On the other hand, orientations recalculated in this thesis align with the orthophotos. Overall, south, south-west, and south-east orientations were prioritized during the installation of PV systems

[7] New roofs dataset with recalculated azimuths: `https://github.com/amilisa/solar-potential-analysis/blob/main/data/roofs/recalculated_roofs.csv`

on real rooftops. At the same time, the northern sections were entirely excluded from installations in all sampled buildings. This is basically an obvious outcome since rooftops with such orientation will hardly produce energy in the northern hemisphere.

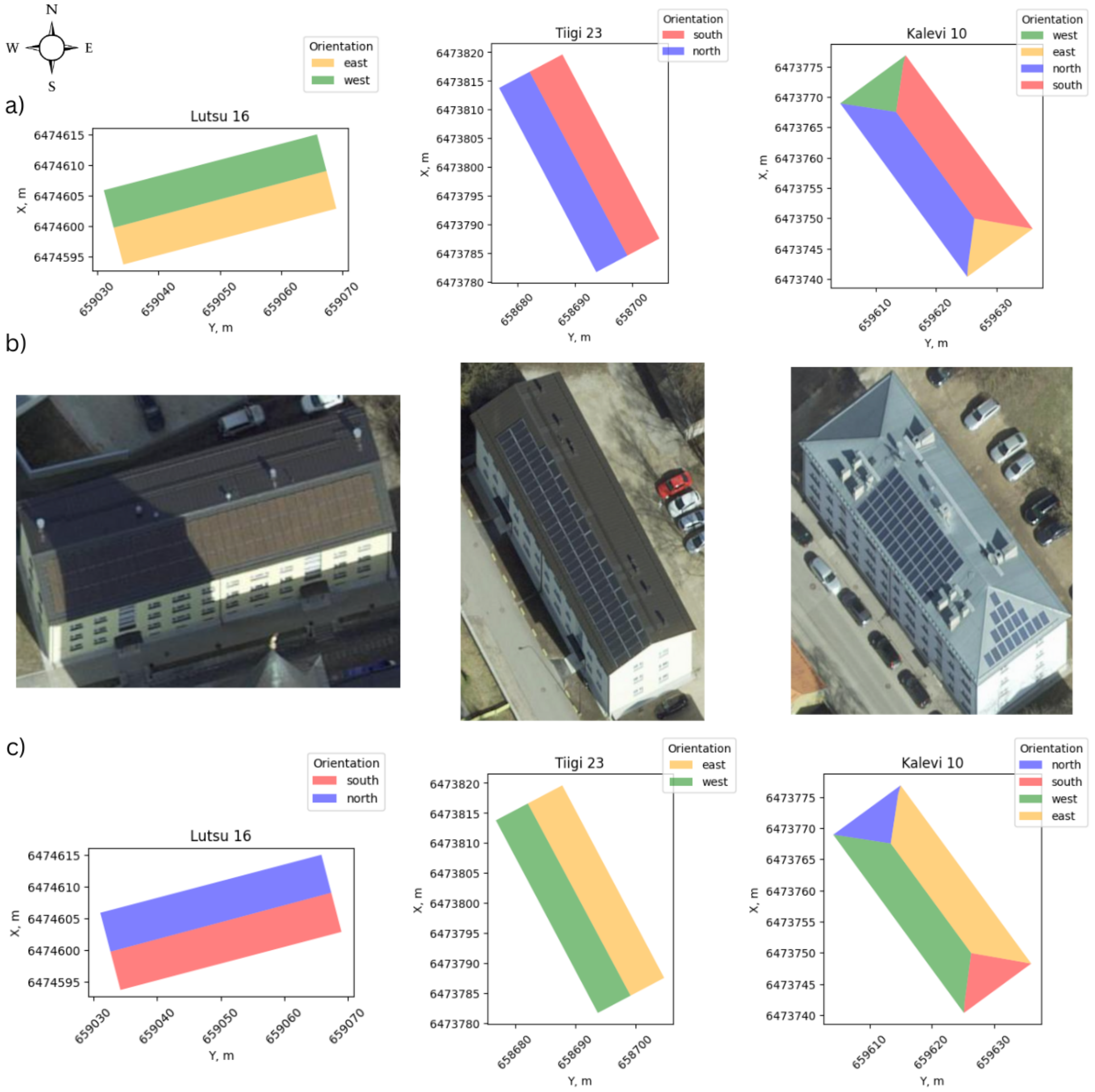


Figure 7. Examples of the plotted rooftops and their orthophotos: a) rooftops with initially assigned orientation; b) orthophotos; c) rooftops with recalculated orientation.

After the identification of the roof parts with installed solar panels, the production output for SmartEnCity buildings was recalculated based on the corrected azimuths, while keeping other parameters, such as panels efficiency and loss, the same as in the previous work. Three types of orientation change were observed: east orientations became south, north became west, and south became east. On average, these orientation changes resulted in the production increase of nearly 12% for east to south case and 35% for north to west. However, for south to east change, there was a decrease of around 27%. The average difference between old and new energy output was about 22%. It is important to note that flat roofs are not assigned with any orientation since the azimuth of horizontal surfaces is 0° and solar panels placement is optimized be PVGIS API. Therefore, flat roofs were not influenced by azimuth recalculations and were excluded from this comparison. In general, the increase or decrease in production highly depends on the extent of the angle change. For instance, in the data analysed, most changes from east to south led to only a slight energy output growth as the angles remained within the south-east direction range (considering a more granular orientation classification). During this process, it was also discovered that a decimal value of 0.14 was previously passed as system losses to the PVGIS API, instead of a percentage. This does not align with both the API documentation [4] and the losses stated in Bohdan Romashchenko's thesis. Such a small loss value is highly unlikely and moreover, it results in overly optimistic solar production output.

Regarding the sizes of roof surfaces available for PV installation, approximately 58.6% of the roof area, on average, was allocated for the placement of solar panels. The estimated in previous work areas, which were assumed to account for 90% of the total roof size, turned out to be 60% larger than the actual ones. Flat and tilted roofs were also considered separately due to potential differences in the specifics of solar panel installation or available area. However, no significant difference was found, with PV systems occupying approximately 55% and 60% on average of flat and tilted roofs, respectively. Additionally, it is worth noting that it is not necessary to use a specific roof area during the energy production estimation step. Instead, PVGIS API calls can be performed for just one square meter for every roof and then multiplied by the desired roof size to obtain the total potential. The resulting energy output has been proven to be the same, but the latter approach is more convenient as the requests can be run once, while the energy production for the entire roof can be easily calculated several times.

To conclude this subsection of the chapter, the analysis of real PV installations has provided valuable insights into developing a filtering strategy for rooftops' parts and determining the average available area. Firstly, north-oriented roofs can be entirely excluded during solar potential estimation, while south, south-west, and south-east orientations should be prioritized. Secondly, for more accurate energy output estimations, the area allocated for PV systems should be set at approximately 60%. Alternatively, calculating the solar potential per square meter could be a better option. This would allow for greater flexibility in determining the extent of surface coverage with panels on different rooftops.

3.3 Recalculating solar energy production potential

The validation process and analysis of rooftops with currently installed solar panels highlighted the need to improve the estimations of Tartu's solar potential, which were previously presented in Bohdan Romashchenko's thesis. In the current study, the buildings were first mapped with their metadata, such as construction year, total floor area, and use purposes. This comprehensive data allows for a more precise analysis of the solar energy production potential, enabling informed decision-making on prioritizing specific construction types for optimal energy output. Before running PVGIS API requests, the data was also cleaned and filtered as real-life installations only considered promising rooftops for PV systems. Additionally, a caching mechanism was implemented to speed up the requests processing making it easier not only to rerun estimations but also to handle the computations for much bigger cities than Tartu.

3.3.1 Mapping buildings with metadata

Prior to the main data cleaning and filtering process, buildings were mapped with metadata to identify their use purposes and exclude garages from the dataset. While some specific cases may have solar panels installed on garage roofs, such occurrences were considered rare due to a lack of information found during the literature review. Additionally, many old garages in Tartu have roofs unsuitable for solar panel installations. Thus, to provide a more realistic estimation of Tartu's solar potential, only buildings with other usage types were considered.

To map buildings with their metadata, this study utilized Buildings Actual Data API[8] provided by E-ehituse platform[9]. One of the endpoints of this API returns buildings data based on the National Building Register code or, in Estonian, Riiklik ehitisregister (EHR). EHR codes were obtained from Tartu's 3D data which is freely accessible on the Estonian Land Board Geoportal[10]. However, not all the buildings contained the EHR code in the 3D data, thus only 13954 buildings out of 20729 were mapped with metadata.

The main metadata fields extracted from the API response included name, construction year, total floor area, and use purposes along with their corresponding areas. The use purposes had a fairly high granularity with 139 distinct values, which was not convenient for the further application of data. First of all, this granularity makes it difficult to work with the dataset, because in cases where a single building has multiple uses, there will be multiple rows in the dataset essentially representing the same building. In addition, it will not be possible to effectively group buildings into meaningful and more generalised

[8] Buildings Actual Data API: `https://swaggerui.ehr.ee/ehitise_kehtivate_andmete_teenus#/Ehitised/get_buildingData`. [28.07.2023]

[9] E-ehituse platform: `https://developer.e-ehitus.ee/`. [28.07.2023]

[10] Estonian Land Board Geoportal: `https://geoportaal.maaamet.ee/eng/Download-3D-data-p837.html`. [28.07.20233]

classes. To address this, the author manually classified the distinct purposes into nine groups. Table 4 represents specific use purpose groups along with the examples of the buildings types included in each category.

Table 4. Examples of buildings included in each use purpose class.

Use Purpose Class	Examples of Buildings
Residential	Houses, Apartments, Hotels
Assembly	Theaters, Restaurants, Sport Halls
Business	Offices, Shops, Banks
Educational	Schools, Universities
Institutional	Government Buildings, Hospitals
Industrial	Factories, Plants
Mercantile	Retail Stores, Malls
Storage	Warehouses, Garages
Other	Buildings, not suitable for other classes

This classification was based on the descriptions of each building class formulated in Civil Engineering [1] and might require refinement as this thesis did not focus on defining the best suitable classification. Consequently, columns representing the classes were added to the dataset, and for each building, the ratio of the area allocated to different purposes was calculated. For example, if a building has the total floor area of 500m^2, with 350m^2 used for residential purposes and 150m^2 allocated for business, the corresponding columns would contain 0.7 and 0.3, respectively. All other columns would contain 0. This approach provides a better understanding of building uses and is inspired by the works of Constantine E. Kontokosta and Christopher Tull [17] described in the chapter 2.

3.3.2 Data cleaning

As the first step of the mapped buildings data cleaning process, constructions having a 'garage' use purpose were excluded, aligning with the explanations provided at the beginning of the Subsection 3.3.1. Following the removal of garage-type buildings, the dataset was further explored, revealing that some structures had exceedingly small total roof area and total floor area. Table 5 demonstrates the general statistics on these columns, which summarizes the central tendency, dispersion and shape of data distribution, excluding empty values. From the table, it is evident that the data contains outliers with the minimal total roof and floor areas recorded as 2m^2 and 0m^2, respectively. To achieve a more realistic solar energy potential estimation, it is necessary to filter out these very small roofs, as they are more likely related to auxiliary constructions and unsuitable for PV systems. For solving this problem, a filtering metric was applied by using the average value of the garage roof in Tartu as a threshold. Consequently, all buildings with a total roof area lower than this threshold, which is 28m^2, were removed from the dataset. The total floor area was filtered using the same threshold, retaining only buildings with a floor area at least equal to the average garage size. This was done to somehow filter out suspiciously small buildings from the dataset as they would hardly be representative in this case. By implementing this approach, the study eliminates potential outliers and focuses on constructions with more suitable roof areas for potential solar panel installations, ensuring a more accurate assessment of Tartu's solar energy capacity.

Table 5. Total roof area and total floor area columns statistics.

Type	Total Roof Area	Total Floor Area
count	20729	13677
mean	243.14	566.35
std	698.53	1814.05
min	2.02	0
25%	48.03	102.80
50%	134.92	191.90
75%	218.40	333.70
max	36618.55	109398.10

Cleaning the building-level data was a preliminary step facilitating the cleaning of rooftops dataset described in the Subsection 3.1.2. The process involved several key steps to ensure the dataset's suitability for solar energy potential estimation. Firstly, a list of building IDs was retrieved, allowing us to focus solely on the relevant parts of each rooftop. Next, the parts with a tilt angle greater than 60° were excluded from the dataset, accounting for approximately 3.3% of all rooftop sections. Based on the paper by Teolan Tomson [21], such parts are too steep to be suitable for PV installations at the latitude of Estonia. Drawing insights from Subsection 3.2, which encompassed an analysis of actual PV installations in Tartu's SmartEnCity buildings, the roofs were further filtered based on their orientation. Specifically, sections more north-oriented, namely north, north-west and north-east, were excluded from consideration. It is not practically beneficial to install solar panels on such roofs as their energy output would be significantly limited in the nothern hemisphere. Additionally, any roof sections that could not accommodate at least two average-size solar panels (approximately 4m^2 in total) and were not in contact with larger section of the same orientation were removed from the dataset. This step was taken to exclude sections that might be occupied by ventilation systems or other structures essential for building functionality, which would not be suitable for solar panel placement. Overall, after completing the comprehensive data cleaning process, approximately 42% of roof sections and nearly 30% of total roof area were filtered out, leaving 45993 sections and 16757 buildings in the dataset. This approach presents a strategy for selecting roof parts to identify more suitable rooftops for potential solar panel installations and provide a more reliable assessment of Tartu's solar energy capacity.

3.3.3 Energy production recalculation

In this subsection, the recalculation of Tartu's solar energy production potential is described, which involved implementing several optimizations to enhance the efficiency of the process. The primary goal was to improve the estimations of solar energy production potential for rooftops, while reducing the computation time. The code for sending requests to the PVGIS API was already implemented to run asynchronously in the previous work. However, due to the rate limiting of 30 requests per second imposed by the API, the implementation required the introduction of timeouts to manage the requests effectively. The non-filtered data and a large number of requests resulted in a considerable computation time, taking 65 minutes to complete the calculations. While this calculation time might not present a considerable concern for Tartu, it will hold much greater importance for cities that are several times larger.

To address this, a caching mechanism was introduced. The idea behind the use of cache was to reduce the number of duplicate requests to the API by storing and reusing previously obtained estimations. The caching system was based on a combination of tilt and azimuth values. When requesting the estimation of solar energy production for a specific combination of tilt and azimuth, the result is stored in cache. For subsequent buildings with the similar tilt and azimuth, the estimation is retrieved from the cache, eliminating the need for redundant API requests. As the author employed Python for working with data and analysing it, the asynchronous code for sending requests and parsing responses was written in this language, making it easy to incorporate further analysis. Caching mechanism was implemented with the use of dictionary which is a Python built-in data structure based on a hash table.

Before implementing the caching mechanism, certain preparations were made to ensure the validity of the approach. Initially, a thorough analysis was conducted to determine whether using a constant latitude and longitude for all buildings in Tartu would introduce significant inaccuracies in the estimations. To assess this, five points were chosen from the north, south, west, east, and central regions of Tartu. Solar energy production per square meter was calculated for each point. The average difference between these points was found to be a mere 0.1%, indicating that using constant latitude and longitude for all buildings would not compromise the accuracy of the estimations. Moreover, to minimize the variation between tilt and azimuth values, they were rounded to integers. This rounding process resulted in fewer distinct combinations of tilt and azimuth values, reducing their number from 44435 to 11211. Thus, the caching mechanism would be more efficient, and the number of unique requests to the API would be significantly reduced.

During the recalculation process, the PVGIS API's loss parameter was set to 14%.

This number, determined in the Bohdan Romashchenko's work [20], represents an average total PV system loss due to factors such as shading, alternating current inverting, and other factors commonly encountered in real-world installations. The panels efficiency was set to 18% which is an average value observed in most commonly used solar panels, according to the web resource by University of Michigan [11]. Additionally, to streamline and standardize the estimation process, a crucial improvement was made by performing the calculations for one square meter across the entire dataset. This approach allowed for a uniform and consistent evaluation of solar energy production potential for each building, independent of their varying sizes and roof areas. Previously, the calculations were done for the total areas of roof parts, which introduced complexities and variations due to the different sizes of rooftop sections. However, by switching to calculations per one square meter, the process became more straightforward and easier to interpret. This standardization not only facilitated the implementation of the caching mechanism but also made the results directly comparable making data processing more efficient.

Finally, after running the requests and obtaining the responses from the PVGIS API, the output data were parsed and mapped with the roofs' metadata based on the corresponding combination of tilt and azimuth. This ensured that the recalculated solar energy production potential[11] was accurately associated with the correct roofs. Overall, data cleaning and cache utilization decreased the computation time to 13 minutes (against 65 minutes in the previous work), greatly accelerating the process and making it easier to apply the calculations not only to Tartu but also to larger cities in the future. In addition, modifying this approach by selecting wider ranges where the combination of slope and azimuth can be considered similar to the cached one could further speed up the process, but at the cost of reduced accuracy.

3.3.4 Validating recalculated production

The present study conducted a comparative analysis between the new energy production estimations[12] per square meter and the actual production data from the available SmartEnCity buildings. This comparison was carried out using a methodology similar to that described in Subsection 3.1.3. The recalculated production values resulted in a Mean Absolute Percentage Error (MAPE) of 11.82%, signifying a considerable improvement when compared to the previous research. The obtained MAPE values for each season are presented in Table 6. As anticipated, winter months exhibited the largest error, with minimal improvement observed. Conversely, the accuracy for other months, such as Spring and Summer, demonstrated better results, with errors of approximately 13%. It is

[11] Recalculated production for roofs: `https://github.com/amilisa/solar-potential-analysis/blob/main/data/roofs/estimated_production.csv`

[12] Recalculated production for buildings: `https://github.com/amilisa/solar-potential-analysis/blob/main/data/buildings/new_estimated_prod_by_building.csv`

crucial to acknowledge that the new estimations correspond to the year 2023, whereas the actual production data relates to 2022. Consequently, there might be an average year-to-year variation[13] of nearly 5% for the buildings under examination.

Table 6. Mean Absolute Percentage Error for each season in recalculated estimations.

Season	MAPE, %
Winter	118.05
Spring	12.37
Summer	13.01
Autumn	18.74

Despite the progress achieved, a significant error persists for Autumn months. Moreover, a noteworthy factor is that the previous work's calculations were based on incorrectly defined azimuths, making a direct comparison of the errors somewhat unfair. As was noted in the Subsection 3.2, depending on the extent of the orientation change, the energy production can increase or decrease by 22% on average. With correct azimuths, previous work results could change greatly, leading to different errors. Thus, to evaluate the accuracy of the new estimations, it is more appropriate to focus on the proximity of the obtained estimations to the real values, rather than conducting a direct comparison with the previous work.

3.4 Using text-generating language model

This thesis used text-generating language model, namely ChatGPT[14], for reviewing the content and improving sentences structure. In addition, DeepL Write tool [15] was used to look up for synonyms and choose words that are more suitable for a particular description. Although text-generating models were used across the whole thesis text for clarity enhancement and better use of the English language, the content itself was created solely by the author.

[13] The year-to-year variation is one of the output fields from the API response which is not present in the previous work.

[14] `https://chat.openai.com/`

[15] `https://www.deepl.com/write`

4 Energy balance

This chapter provides a more detailed examination of the existing methodologies for energy consumption prediction aiming to answer the second part of research questions. Additionally, the available data on Tartu's electricity consumption is described, and energy balance evaluation is performed based on this data.

4.1 Approaches to energy consumption prediction

Understanding buildings' energy consumption is crucial for establishing an accurate energy balance and enabling sustainable urban development. As it was explained in the chapter 2, accessing energy consumption data at a city-scale is hardly possible due to privacy concerns and confidentiality requirements. To address this limitation, predictive models can be used to offer a practical solution by estimating energy usage based on aggregated data and various contextual factors. These models allow data-driven decisions without compromising individual-level data privacy. In this thesis, various modelling approaches have been investigated in the field of energy forecasting, ranging from sophisticated deep learning methods to more traditional statistical methods. This section will focus on regression models as they have been employed in many studies and have shown to be perspective methods for predicting energy consumption in buildings [13]. Moreover, while deep learning methods have shown promise in various domains, regression models have gained attention due to their simplicity, interpretability, and acceptable accuracy in this area of research. By analysing the relevant literature, the author seeks to lay the groundwork for determining a suitable model for predicting electrical energy in Tartu in the future.

The paper by Alessio Mastrucci *et.al* [18], introduced in the chapter 2, presents a methodology for predicting the energy consumption of dwellings based on the use of a multiple linear regression and downscaling technique. The main idea of the methodology is to employ a regression model for downscaling energy consumption data from post-code area level to individual dwellings level. The methodology was built around a case study of Rotterdam city with 993 post-code areas and nearly 300000 dwellings. The residential buildings were classified into six different categories in accordance with the national classification. Additionally, different characteristics of buildings, including type, construction year, floor surface, and number of occupants were retrieved for every address. Aggregated and processed metadata was then mapped with the zip-code level electricity consumption values requested directly from the city. Table 7 describes the post-code level input features for the model. The proposed by Alessio Mastrucci *et.al* approach involved regression equations at the post-code area level and the Ordinary Least Squares (OLS) method was employed to fit the model. The model's formulation for electricity consumption is as follows:

$$y = \beta_0 + x_{floor} \cdot \beta_{floor} + x_{people} \cdot \beta_{people} + \sum_{i=1}^{5} (x_{type,i} \cdot \beta_{type,i}) + \epsilon \tag{1}$$

After fitting, the authors first calculated prediction error as the Mean Absolute Percentage Error. Next, they carefully validated their model through bootstrapping [10], a process that assesses the accuracy of the predictive model using the available dataset. In this technique, the original input data are randomly divided into training and test sets. The former is used to fit the model and the latter is used to compare measured values with predicted values. The model accuracy was evaluated by 'optimism' index which defines the difference in the values of coefficient of determination, R^2, and Mean Squared Error (MSE) for training and test sets [18].

Table 7. Input data at post-code area level in the work by Alessio Mastrucci *et.al* [18].

Data	Variable	Unit
Average floor area per dwelling	x_{floor}	m^2
Share of dwellings per type	$x_{type,i}$	$\%$
Number of occupants	x_{people}	n
Yearly average electricity consumption per dwelling	y	kWh

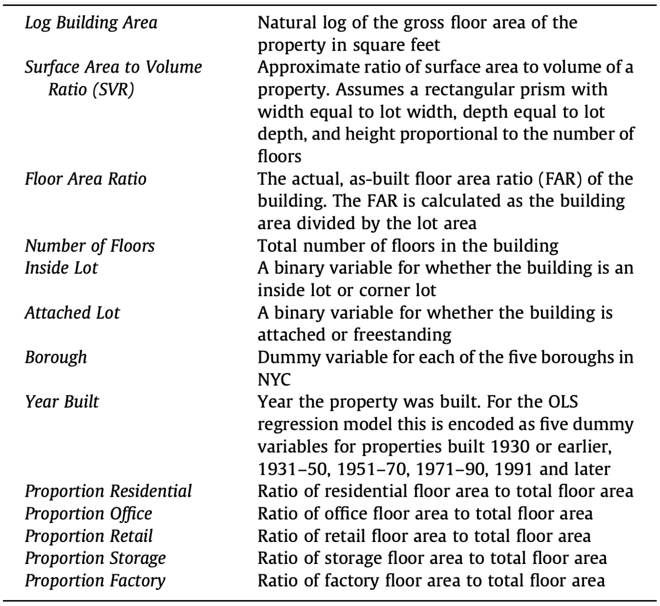
As the output of linear regression model, the coefficients $\hat{\beta}$ were calculated and subsequently used to downscale electricity consumption. For every dwelling, given the dwelling type t, floor surface x_{floor}, and the number of occupants x_{people}, the electricity consumption $\bar{y}$ was predicted in accordance with the following formula:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_{type,t} + x_{floor} \cdot \hat{\beta}_{floor} + x_{people} \cdot \hat{\beta}_{people} \tag{2}$$

where $\hat{\beta}_0$, $\hat{\beta}_{type,j}$, $\hat{\beta}_{floor}$ and $\hat{\beta}_{people}$ are the estimated regression coefficients [18]. The statistical analysis showed that the regression model was fairly accurate in predicting electricity consumption, which was identified by low 'optimism' value. The R^2 of 0.817 demonstrated that approximately 81.7% of the variability in electricity consumption across the city is accounted for and explained by the combination of building characteristics used in the regression model. The Mean Absolute Percentage Error is found to be 9%, indicating good prediction accuracy. Overall, the methodology created by Alessio Mastrucci *et.al* provides an effective approach for electricity consumption prediction at the city scale without compromising data privacy. Although the work only predicts annual electricity consumption for residential buildings, the model can be modified in terms of variables introduced or used to predict monthly energy consumption.

In the paper by Constantine E. Kontokosta and Christopher Tull [17], discussed in chapter 2, an alternative methodology for energy consumption prediction in New York

City (NYC) was employed. Although the authors also used regression models, they approached the problem in a different manner. They leveraged data from two primary sources: New York City's energy disclosure policies and confidential and more detailed version of that data provided by the NYC Mayor's Office of Sustainability. This allowed them to directly train their model using the relevant data. Before the training stage, the authors matched various building characteristics with consumption data. To optimize their model's performance, they employed a stepwise selection process to identify the most relevant features minimizing prediction errors. The selected variables are illustrated by Figure 8. It be clearly seen that as the training data was more granular compared to one from the work by Alessio Mastrucci *et.al*, predictors were more closely associated with individual buildings. Furthermore, Kontokosta and Tull's research went beyond focusing solely on residential stock. They considered multiple types of buildings and introduced the ratios of areas allocated to various use purposes, offering a city-scale understanding of energy consumption patterns. The paper also compares three models, namely Ordinary Least Squares regression model (OLS), Support Vector Regression, and Random Forest. The results showed that OLS ranked first overall in term of prediction performance.



Log Building Area	Natural log of the gross floor area of the property in square feet
Surface Area to Volume Ratio (SVR)	Approximate ratio of surface area to volume of a property. Assumes a rectangular prism with width equal to lot width, depth equal to lot depth, and height proportional to the number of floors
Floor Area Ratio	The actual, as-built floor area ratio (FAR) of the building. The FAR is calculated as the building area divided by the lot area
Number of Floors	Total number of floors in the building
Inside Lot	A binary variable for whether the building is an inside lot or corner lot
Attached Lot	A binary variable for whether the building is attached or freestanding
Borough	Dummy variable for each of the five boroughs in NYC
Year Built	Year the property was built. For the OLS regression model this is encoded as five dummy variables for properties built 1930 or earlier, 1931–50, 1951–70, 1971–90, 1991 and later
Proportion Residential	Ratio of residential floor area to total floor area
Proportion Office	Ratio of office floor area to total floor area
Proportion Retail	Ratio of retail floor area to total floor area
Proportion Storage	Ratio of storage floor area to total floor area
Proportion Factory	Ratio of factory floor area to total floor area

Figure 8. Features selected for prediction models in the work by Constantine E. Kontokosta and Christopher Tull [17].

In conclusion, energy consumption prediction remains a relevant and challenging task, considering data availability constraints due to privacy concerns. The research highlights the suitability of linear regression in identifying correlations between energy consumption and its predictors. While granular building-level energy consumption data facilitates straightforward model training, alternative approaches such as downscaling from the aggregated postcode level proposed by Alessio Mastrucci *et.al* offer a promising methodology that preserves data confidentiality. Adapting the multiple linear regression model for all buildings represents a relevant extension of their methodology. One approach could be to replace the shares of dwelling types with the ratios of various use purposes of buildings. The use purpose of the building might provide a rough estimate of the number of occupants, but it may not always accurately reflect the actual occupancy, especially for buildings with unique use patterns. In cases where precise data on the number of occupants is not available, caution should be exercised when assuming the number of occupants based solely on the building use purpose. Including other relevant predictors or exploring alternative methods for estimating the number of occupants may be necessary to improve prediction accuracy. To predict monthly energy consumption, the inclusion of additional features, such as average monthly temperature, aligns with insights from the paper by Tao Hong *et. al* [15], emphasizing the influence of temperature seasonality on electricity consumption. Overall, when extending the Mastrucci *et.al* methodology to predict consumption across various building types, it is important to explore additional features such as building-specific use patterns, energy efficiency measures, weather conditions, building age, or construction type.

4.2 Available data on the energy consumption of buildings in Tartu

With the help of this thesis supervisor and Tartu City Government, a total of 315 *csv* were acquired from Cumulocity platform. These files contained data on electricity consumption over three years, from 2020 to 2022. Each file included electricity consumption for 12 months with timestamps, consumption values, and type of series. However, nearly 43% of all the files were found empty, with 2020 having the highest number of empty files. The percentages of empty files for each year separately is illustrated by Figure 9. The total number of buildings with existing data was equal to 82. Overall, after a combined computational and manual analysis, it became evident that the data quality was questionable, necessitating preliminary processing before evaluating consumption.

The initial data preprocessing involved removing duplicated rows and eliminating negative consumption values. Additionally, depending on the series type, electricity consumption values were either cumulative or not, thus requiring separate processing. Cumulative values sometimes displayed inconsistent trends, with some months starting at 0 and gradually increasing, while others showed periodic fluctuations, likely indicating system counter failures. To calculate monthly consumption for cumulative series, the maximum consumption value was chosen for each month. For non-cumulative series, monthly consumption was aggregated by summing values for months with at least 25 days of measurements, approximating total electricity usage. After performing monthly calculations, some data points were found to be 0 and were subsequently excluded from

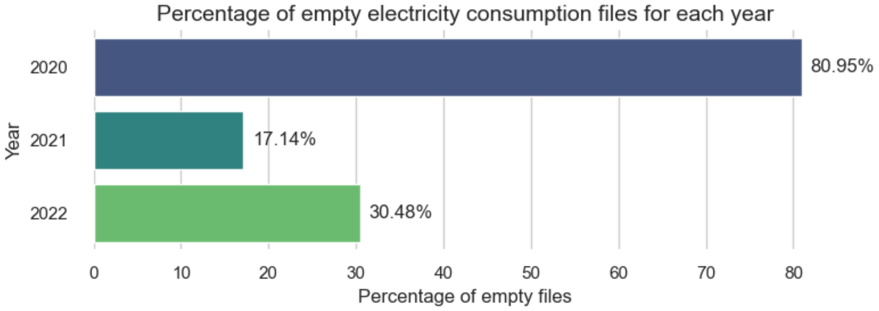


Figure 9. Percentage of empty electricity consumption files for each of three years.

the dataset.

Following the data processing, a total of 67 constructions remained which constitute only about 0.4% of Tartu's building stock (considering the number of filtered constructions from the Subsection 3.3.2). Next, the buildings were mapped with their metadata obtained in the Subsection 3.3.1 based on building IDs. The analysis revealed minimal variation in the types of building area use purposes, with the majority being either partially or completely residential. Other types were underrepresented in the dataset, and around half of the constructions lacked available metadata. Table 8 provides an overview of use purposes and the number of buildings associated with each category.

Table 8. Use purposes and the number of buildings associated with each category.

Use purpose	Buildings number
Residential	30
Assembly	1
Business	8
Educational	1
Industrial	0
Institutional	1
Mercantile	7
Storage	0
Other	0

Overall, the acquired data represents only a small fraction of Tartu's building stock. Given its limited quantity and quality, the data is unsuitable for predicting electricity consumption for city-scale energy balance evaluation. The question of how much data is needed to train a model is quite debatable and depends on both the task itself and the number of features used. However, when it comes to data quality, one of the most important factors is the representativeness of the observations. This means that there must be adequate coverage of the different classes or categories. Otherwise, sampling bias or lack of variance and fluctuations in the training data will lead to inaccurate predictions [14].

4.3 Energy balance evaluation based on available data

Since there was a subset of buildings with accessible data on energy consumption, the author chose to calculate the energy balance for them. Moreover, by obtaining the aggregated electricity consumption value for the entire city, a comparison was made between the estimated solar energy production and the overall electricity consumption. The total electricity consumption for Tartu in 2022 amounted to 465831 MWh. To estimate the annual solar energy production, calculations were performed for various percentages of roof area coverage with photovoltaic (PV) systems. Figure 10 illustrates the potential extent to which solar energy production could cover total Tartu's electricity consumption if PV systems were installed on promising rooftop sections of 16757 buildings.

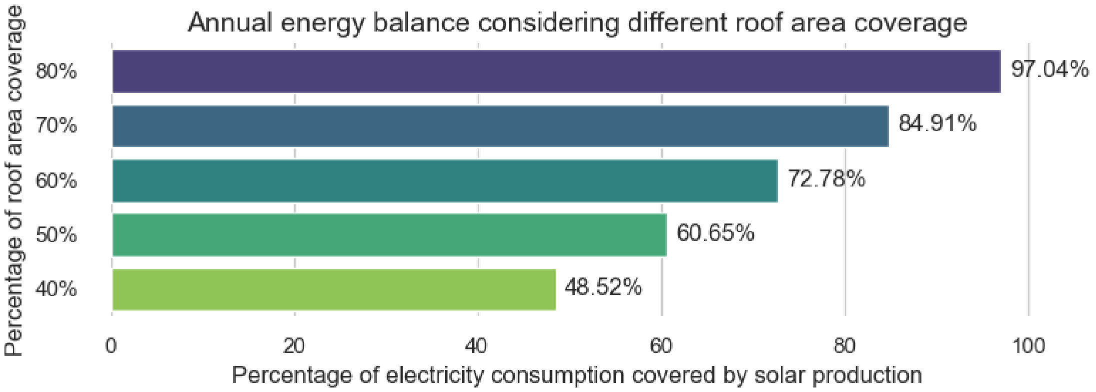


Figure 10. Overall Tartu's potential annual energy balance for 2022 for different values of PV allocated area.

As the roof area covered with solar panels increases, the percentage of electricity consumption covered by production growth by approximately 12%. At 80% roof area coverage, solar production covers around 97% of electricity consumption. This indicates that the solar panels could generate almost all of the electricity needed to meet the consumption demands, leaving only a small portion to be sourced from other means. Even with 50% or 60% of roof sections coverage, substantial amount of electricity can be produced by solar panels. Due to the year-to-year variations or a different roof area availability, these numbers can be lower. However, there might be also periods of surplus energy production, where the solar panels generate more electricity than needed for consumption. This opens up opportunities to store or export excess energy to the grid, contributing to a more sustainable energy ecosystem. This preliminary analysis serves as an initial exploration of the energy balance dynamics in the city, offering valuable insights into the feasibility of utilizing solar energy sources to offset electricity consumption.

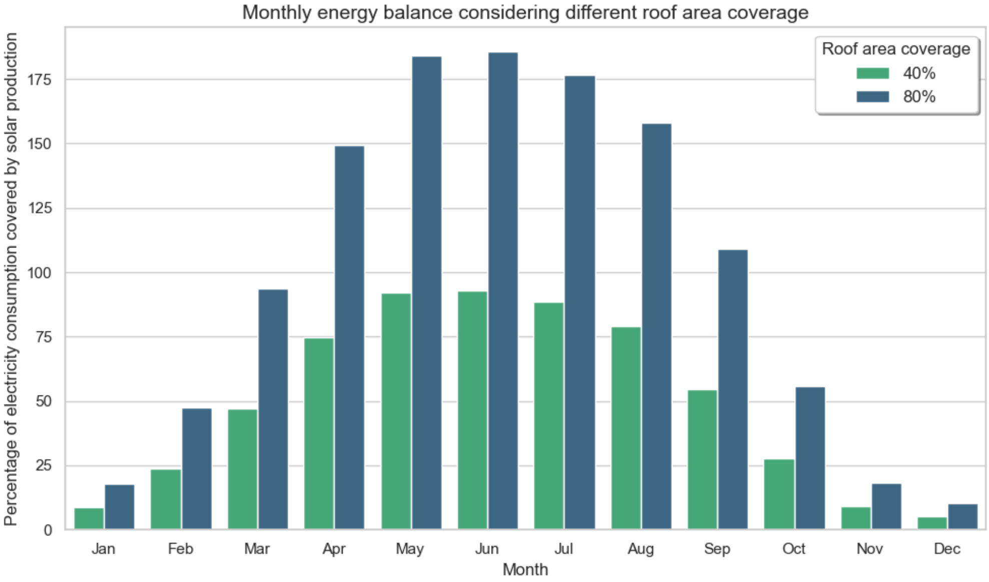


Figure 11. Overall Tartu's potential monthly energy balance for 2021 for different values of PV allocated area.

The total Tartu's energy balance for each month of 2021, for which data was available, was also reviewed. Figure 11 compares the potential monthly energy balance for 40% and 80% of roof area that can be covered by solar panels. There is a clear seasonal pattern, with both ratios exhibiting higher values in the middle months, particularly in May, June and July, and lower values in the winter months, notably December and January. In the summer months, a 40% roof coverage fulfills almost the entire electricity demand, while an 80% coverage even leads to surplus energy production. However, during November, December, and January, the energy balance drops to below 25%, irrespective of the roof coverage size. This suggests a substantial reliance on solar energy during the summer, but additional measures might be needed to ensure sufficient energy supply during the winter months. Overall, the data shows a gradual increase in balance from January to June, followed by a gradual decrease from July to December.

It was also possible to compare electricity consumption and solar energy production using actual data for nine SmartEnCity buildings for 2022. As actual production data for December was unavailable, the average between November and January numbers was calculated to fill in NaN values. Thus, an aggregate of 12 months was calculated and

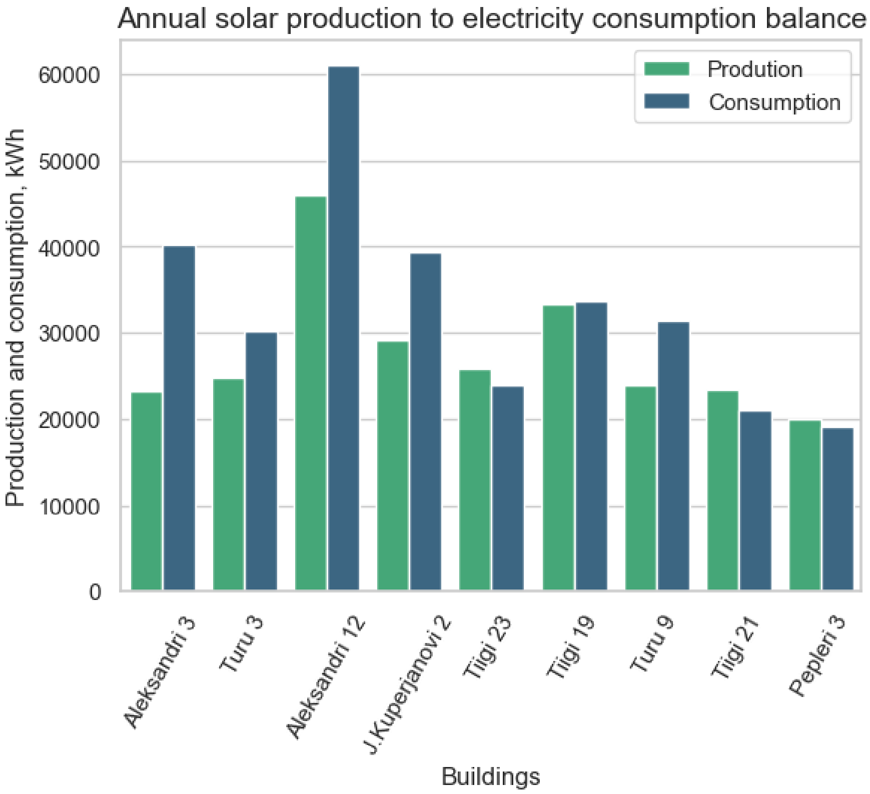


Figure 12. Actual energy balance of SmartEnCity buildings for 2022.

represented in Figure 12 as a bar plot. The results showed that, on average, solar energy covered more than 50% of electricity consumption across all nine buildings. Remarkably, four dwellings, namely Tiigi 23, Tiigi 21, and Pepleri 3, even exhibited slightly higher production levels than their consumption demands. At the same time, for Tiigi 19, the production almost completely covered the consumption, based on the data. It is noteworthy that these outcomes rely on a small number of real data available, and during this study, it was not possible to properly confirm the validity of consumption figures or compare them with some average values. In addition, the balance only took into account electricity consumption calculations, overlooking, for example, energy consumption for heating. Therefore, although the input data for calculating consumption was assumed to be correct, the resulting values should be critically observed, taking into account the lack of validation. Nevertheless, the study underscored the positive impact of solar panel usage on Tartu's energy balance.

5 Results

This section presents the findings from the analysis of the potential solar energy production in Tartu recalculated in this thesis with the use of PVGIS API. First, a descriptive overview of the solar energy data is provided, followed by the distribution analysis of solar energy potential. Subsequently, the available data on construction years, requested from one of the E-ehituse platform APIs (see the subsection 3.3.1), are examined to understand how the suitability of buildings for high solar energy production changes over time. Next, the optimal tilt values yielding the highest production are determined. Additionally, buildings' footprints are visualized as a map colored based on their potential for solar energy production.

To begin with, this analysis aims to evaluate the potential solar energy production in Tartu and provide key insights on the potential within the city to help policymakers and city government make informed decisions in improving Tartu's sustainability. The analysis utilizes data on the solar energy production potential per square meter, estimated and described in detail in the chapter 3 of this thesis. Additionally, the dataset contains various metadata, including building construction years, roof orientations, tilt angles, and districts. Analysing such data plays a pivotal role in understanding the variability, distribution, and factors influencing solar energy potential across different building types and locations in Tartu.

To obtain an overall understanding of the solar energy potential in Tartu, the descriptive statistics for monthly average and annual potential solar production per square meter, depicted by Table 9, is considered. On average, each square meter of solar panel installation could generate approximately 13kWh/m^2 of energy monthly, considering an 18% panel efficiency and 14% losses. The standard deviation of 0.77 indicates moderate variability around the mean value, suggesting that buildings have relatively consistent potential monthly energy production. Regarding annual potential, approximately 155kWh/m^2 can be produced annually on average, with values ranging from a minimum of nearly 117kWh/m^2 to a maximum of around 164kWh/m^2.

In order to provide a broader view of solar potential distribution, the histogram plot for annual solar energy production per square meter is analysed. The histogram is illustrated by the Figure 13 which demonstrates the wide range of potential production values across the city. It reveals that the distribution is characterized by a substantial number of buildings with moderate solar energy production capacity, while instances of low potential are infrequent. Notably, there is a significant number of buildings having high production, with approximately 25% of the observed constructions experiencing the production of 164kWh/m^2. This is explained by the fact that a lot of buildings in Tartu have predominantly flat roofs, which was discovered during the analysis. Figure 14 shows the count of buildings with different predominant orientations, where *none* represents flat roofs. For flat surfaces, the PVGIS API offers optimization of tilt and azimuth angles as solar panels placement on roofs without slope can be easily adjusted.

Table 9. Descriptive statistics for monthly average and annual potential solar production per square meter.

	Annual production, kWh/m^2	Monthly average production, kWh/m^2
count	16757	16757
mean	155.02	12.92
std	9.26	0.77
min	116.82	9.73
25%	149.45	12.45
50%	156.44	13.04
75%	164.10	13.67
max	164.10	13.68

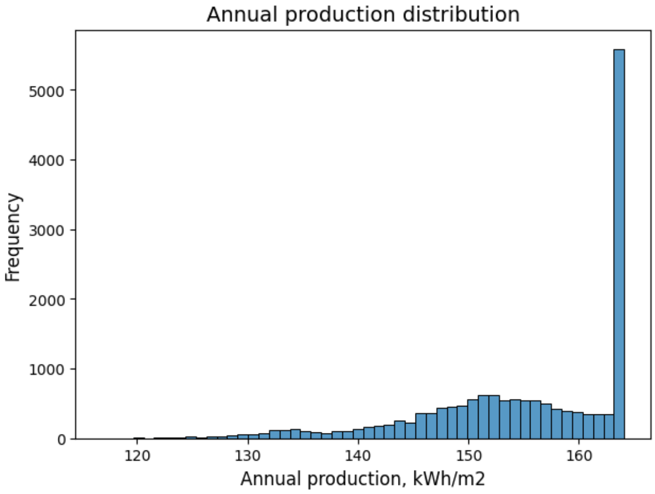


Figure 13. The distribution of annual potential solar energy production per square meter.

Such approach helps to maximize the energy output of installed PV systems, highlighting the importance of rooftop design.

To understand whether the values of Tartu's solar energy potential depend on the year of construction, available metadata data was analysed. Based on the analysis, it was determined that a vast majority of buildings in Tartu were constructed after the 1960s, with the year 1977 witnessing the highest number of building constructions. The average potential solar energy production per square meter for each year of construction from 1960 to 2022 is depicted by the Figure 15. From the graph, it is clear that there is a a noticeable upward trajectory in average potential production over the period, with some fluctuations that can be attributed to differences in rooftop design. It is noteworthy that the buildings constructed around 1983 show the highest peak in potential solar energy

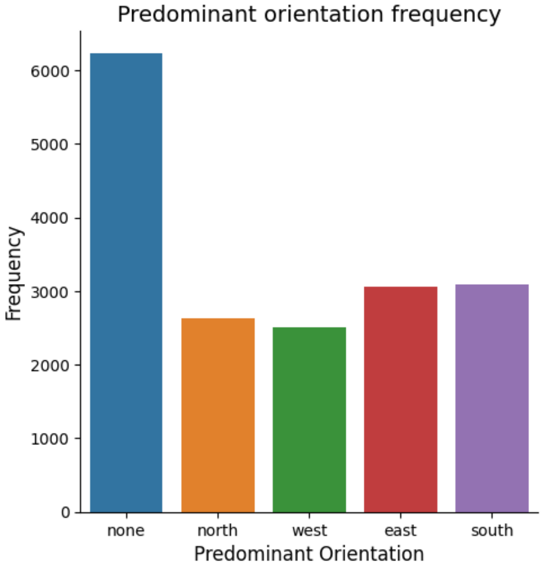


Figure 14. The distribution of predominant orientation of buildings' rooftops where *none* represents flat roofs.

production, indicating favourable rooftop design for potential PV systems installation. In contrast, the years between 1990 and 2000 reflect comparatively lower values. Nevertheless, the rising trend resurfaces from 2000 to 2022, suggesting a resurgence in optimal rooftop conditions for solar installations in newly constructed buildings.

With the potential for optimizing solar panel tilt and orientation, flat roofs emerge as the most promising surfaces for achieving high levels of solar energy production. This insight is further emphasized by the data presented in Figure 16, which illustrates the annual potential solar energy production per square meter for various roof orientations. PVGIS API selects the optimal combination of tilt and orientation that can be applied to flat roofs to optimize energy production estimations for specific location, resulting in maximized production per square meter, as evidenced by the box plot. South orientation, expectedly, also yields substantial output and experiences some fluctuations around the median of approximately 155kWh/m^2. For east and west, the figures are very similar to each other but with more dispersion in comparison to the south orientation. While the dispersion may be attributed to the influence of tilt angles, the fact that a roof can be more south-oriented significantly contributes to these variations for east and west orientations.

Exploring sloped roofs, the conducted investigation revealed that the most optimal tilt angle for achieving energy production levels close to the average per square meter is approximately 40 degrees when combined with a south-oriented roof. Figure 17

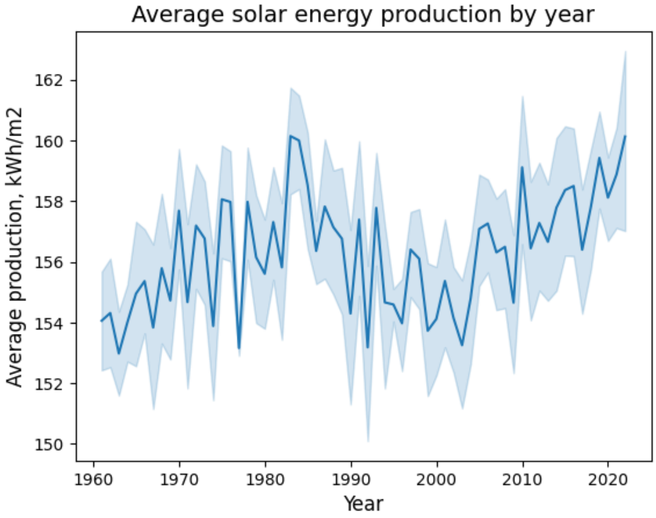


Figure 15. Average potential solar energy production per square meter for different construction years.

demonstrates how average energy production per square meter changes with tilt angle for south, west and east separately. Comparing these findings with the overall average across all roof types, it becomes evident that south-oriented surfaces outperform both west and east orientations. Moreover, for south-oriented roofs with tilt angles ranging from approximately 25 to 50 degrees, the production values are above the overall average. Additionally, tilt between around 38° and 45° shows fluctuations close to the peak production potential. Conversely, east and west orientations consistently yield values below the overall average, with east orientation showing slightly improved performance. Notably, both orientations demonstrate higher energy production with tilt angles around 30-35 degrees.

Spatial insights of solar potential in Tartu were gained by clustering solar energy production using the 25th and 75th quantiles, categorizing buildings into high, medium, and low production clusters. The resulting clusters were visualized on a city map, presented in the Figure18. On this map, the footprints of buildings were color-coded to correspond to their respective production clusters. In this visualization, red color represents buildings with high production capabilities, while green and blue markers denote medium and low production, respectively. This map offers a clear understanding of the spatial distribution of solar energy potential throughout Tartu. Specifically, the northeastern, southern, and southwestern parts of the city, encompassing districts like Annelinn, Ropka, Ropka tööstuse, and Tammelinn, feature a notable concentration of buildings with the potential for higher solar energy production values. This prevalence can be attributed to the prevalence of flat roofs, which can be seen on the actual map, especially for Annelinn. The metadata

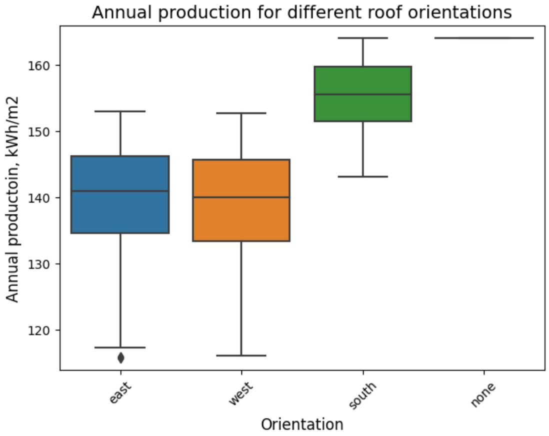


Figure 16. Annual potential solar energy production per square meter for different roof orientations.

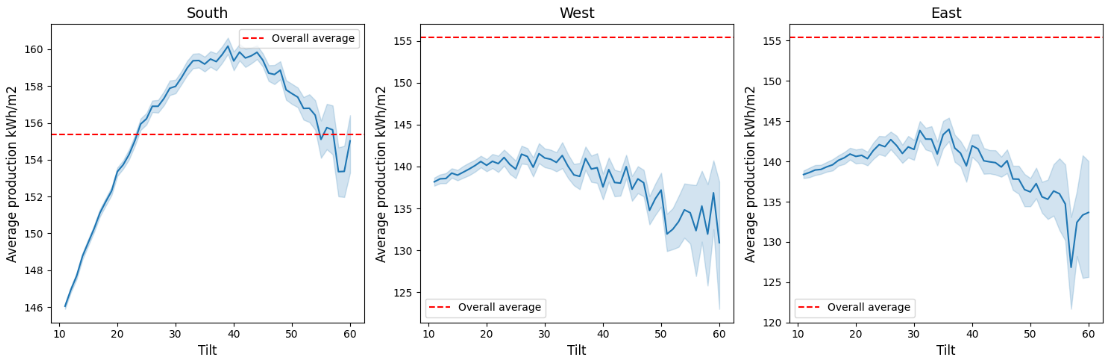


Figure 17. Correlation between tilt and annual production per square meter for south, west, and east.

analysis indicates that these districts predominantly witnessed construction during the 1970s and 1980s, with Ropka tööstuse including active construction periods in the 1990s and 2000s. In contrast, the western districts and city center exhibit medium solar energy potential, as sloped roofs are more common in these areas. The northern region, for example Kvissentali district with a significant number of buildings constructed after the 2000s, showcases a mixture of production clusters. Notably, the Karlova district is characterized by numerous constructions with relatively lower potential for solar energy production.

To conclude, in this comprehensive analysis of the solar energy potential in Tartu, a number of important insights have been drawn that contribute to our understanding

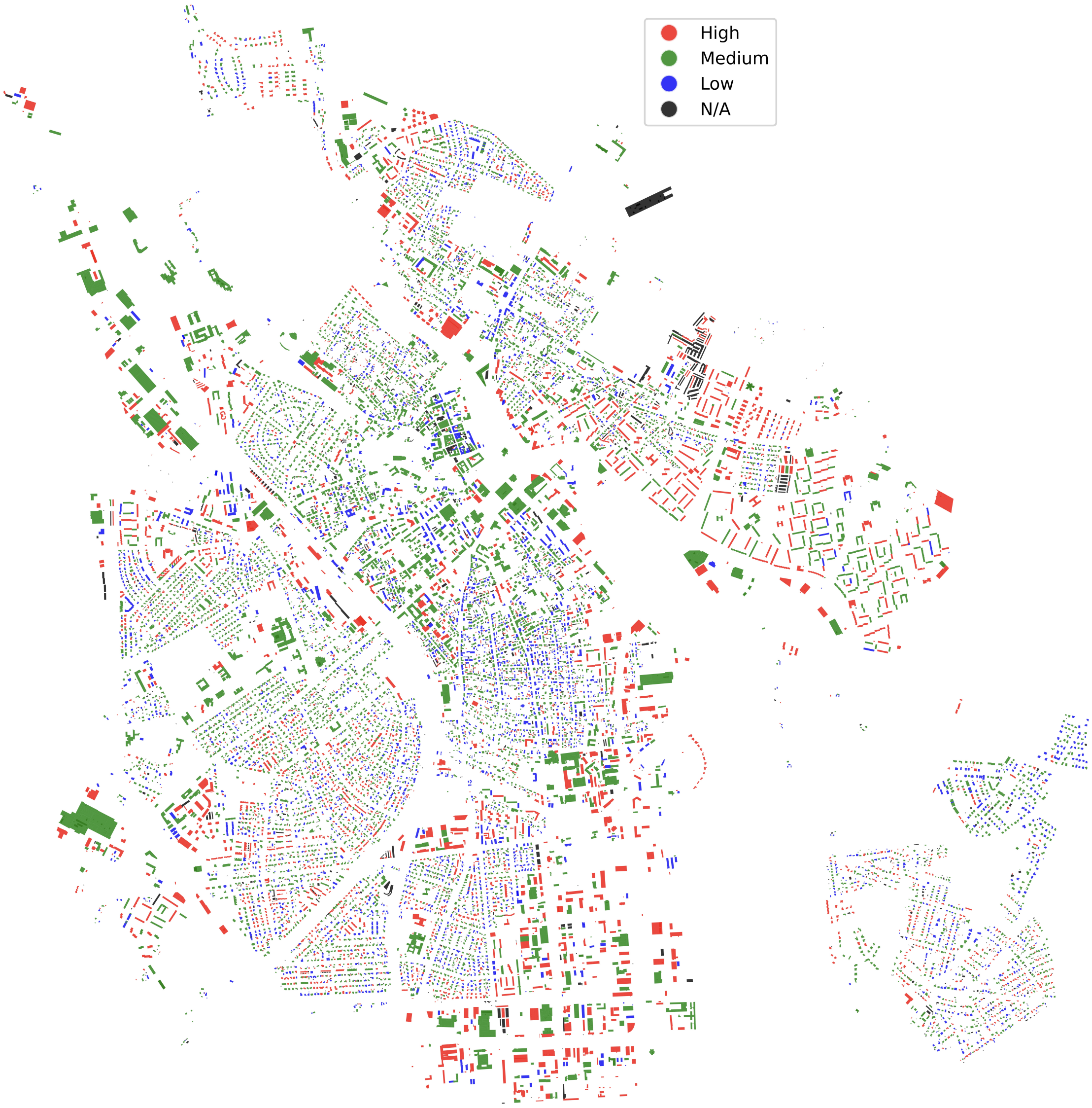


Figure 18. Color-coded buildings' footprints of Tartu representing different levels of solar energy production potential.

of solar energy utilization. First of all, the findings highlighted the significance of roof orientation and tilt angles in influencing solar energy potential. South-oriented roofs with slope around 40° and flat roofs with optimizes solar panels placement showcased the most favorable conditions for maximizing energy production. Overall, the visualization of clustered of energy production values provided a comprehensive overview of solar potential distribution throughout Tartu.

6 Discussion

This chapter answers the research questions discussing the insights gained during the conducted investigations and analysis. Additionally, it outlines the main contributions and provides limitations of this thesis along with the potential directions for future work.

6.1 Answering research questions

The comparison of solar potential estimations made in the previous work with the actual production data from Tartu's SmartEnCity, described in the subsection 3.1, helped to address the research question **RQ1.1**, namely *"How accurate is the solar energy potential estimated in the previous work comparing it to the actual deployments in the city?"*. The findings revealed significant discrepancies between the estimated and actual solar energy production values. The Mean Absolute Percentage Error (MAPE) for the annual production was found to be 23.61%, indicating a notable difference between the predicted and observed values. This discrepancy could be attributed to various factors, such as a mismatch of solar panel efficiency and loss, given as input values for PVGIS API, with the actual parameters. However, the visual representation of roof parts orientations for SmartEnCity buildings, presented in the subsection 3.2, demonstrated that the azimuths were calculated incorrectly, thus negatively contributing to the previous estimations accuracy. Moreover, analysis of real solar energy production data revealed that south, south-west, and south-east orientations were preferred for PV system installations, while northern orientations were excluded due to their limited energy production potential in the northern hemisphere. This showed the need to select more promising roofs for the evaluations to obtain the output closer to reality. As one of the results of this thesis, the azimuths were recalculated, roofs data was cleaned, and new estimations of Tartu's potential solar energy production were performed, with solar panel efficiency and losses of 18% and 14%, respectively. The recalculated production values yielded a reduced MAPE of 11.82%, representing a substantial enhancement over the earlier research. Nonetheless, the fact that the prior calculations were based on inaccurately defined azimuths makes direct error comparison somewhat skewed. As detailed in the subsection 3.2, the extent of orientation adjustments can lead to an average energy production change of 22%. Thus, it is unclear how correctly defined azimuths would influence the results of the previous work. Therefore, for evaluating the accuracy of the new estimations, focusing on the proximity of these estimations to actual values from SmartEnCity buildings is more appropriate. Overall, due to the identified mistakes and necessary corrections, the previous production estimations cannot be considered as precise.

When addressing research question **RQ1.2**, *"What types of buildings (based on their roofs' features) contribute more to the energy potential? Which types of constructions should be prioritised in the future to improve energy potential of the city?"*, a comprehensive examination of the estimated solar energy potential in Tartu revealed various insights. Notably, many buildings exhibited moderate solar energy production capacity, with a considerable portion capable of high production, particularly those with predominantly flat roofs. This can be explained by the fact that constructions with flat roofs facilitate the maximisation of energy production through the possibility of optimising the positioning and tilt of solar panels. The investigation of sloped roofs determined that the optimal combination of slope and azimuth for achieving high levels of energy production is a slope angle ranging from approximately 38 to 45 degrees, paired with a south-facing roof orientation. Additionally, south-oriented roofs with tilt angles between 25° and 50° experience production values higher than the overall average, which is a promising finding. In contrast, west and east orientations tend to yield medium to low output, with the best production observed at slope angles of 30 to 35 degrees. Based on the comprehensive data analysis, it can be concluded that prioritizing flat roofs or south-facing roofs with tilt angles fluctuating around 40 degrees is advisable when designing new buildings. This approach would enhance solar production potential and make a positive contribution to the sustainable development of Tartu.

In relation to research question **RQ1.3**, *"Which city areas contribute most to the energy potential?"*, spatial insights were gained by clustering solar energy production using 25th and 75th quantiles. The resulting clusters were visualized on a city map, indicating distinct concentrations of buildings with high production potential in the northeastern, southern, and southwestern parts of the city. The districts in these regions, especially Annelinn, have a lot of buildings with flat roofs, reflecting the maximum potential estimate of solar energy production. In contrast, areas with sloped roofs, such as the city center and western districts, witnessed medium solar energy potential. The northern region, including districts like Kvissentali, showcased a mixture of production clusters.

The challenges associated with energy consumption data collection and the methodologies used to predict energy consumption were investigated to address the second part of research questions focused on a city-scale energy consumption estimation. Due to privacy constraints, acquiring energy consumption data for all the buildings in the city is highly unlikely. Predictive models offer a solution by estimating energy usage using machine learning approaches. These approaches still require some consumption data for models training to identify correlations between variables and be able to further assess energy usage for all the buildings of a particular city. Even a small amount of data can be difficult to obtain, especially with good quality and enough variation. However, one of the methodologies, investigated in detail in the subsection 4.1, offered the potential to estimate energy consumption without compromising data privacy. Alessio Mastrucci et al. [18] proposed an effective approach for electricity consumption prediction on a city scale using downscaling of aggregated zip-code level consumption to individual buildings with the use of regression model. The used data represented consumption on 993 post-code areas with aggregated metadata which included average floor area per building, number of occupants and a variety of building types. Additionally, in some cases, cities introduce energy disclosure policies for certain buildings as discussed in the paper by Constantine E. Kontokosta and Christopher Tull. The authors introduced an alternative methodology for energy consumption prediction in New York City because they trained the model on more granular data. Nevertheless, in their approach, Kontokosta and Tull also accounted for different building classes and employed regression model for their prediction.

Overall, insights gained from a comprehensive literature review, outlined in chapter 2 and subsection 4.1, helped to answer the research question **RQ2.1**, namely *"What are the suitable models for energy consumption estimation?"*. Through the exploration of various methodologies, including regression and artificial neural networks, it became evident that regression models are promising for energy consumption estimation due to their simplicity, interpretability, and notable accuracy. Such models as multiple linear regression are successfully employed in many works in this domain. They operate on systems of equations containing multiple variables or features that describe the input data, resulting in a parameter set that allows for out-of-sample consumption predictions. In the context of energy consumption prediction, these input variables range from building characteristics and weather conditions to sample consumption values, which together determine the predictive power of regression techniques.

The research question **RQ2.2**, concerning *the possibility to estimate the monthly energy consumption of city buildings based on partial information, which is currently available or can be gathered in short time*, is not easily answered with a simple yes or no response. This is due to the dependence of the answer on the characteristics of the data being available for the estimation. Importantly, partial information in this context can be understood as data on total consumption of different city areas or samples of energy consumption from a limited number of buildings. The research works on these domain proved that the data acquired for training can be partial, but at the same time, it should be representative. The amount of data needed for training is rather related to the problem in question. However, encompassing a range of classes is essential for accurate predictions, enabling the model to learn from diverse categories and generalize effectively.

This, in turn, leads to an answer to the final research question **RQ2.3** focusing on *which data should be collected to perform energy consumption prediction*. Apart from data quality and diversity, the dataset should also incorporate suitable features. Depending on the model used for prediction, the specific requirements for the data to be collected may differ to some extent. Based on the reviewed literature, for a model downscaling monthly zip-code aggregated consumption, the dataset should include variables like floor area, number of occupants or some other information on occupancy patterns, share of buildings per type, post-code level monthly consumption, and average monthly temperature. If consumption prediction is centered around sampled individual buildings, variables like construction year, number of floors, and surface area to volume ratio can be added. Although such requirements lay the groundwork and set the direction, it is essential to recognize that more data and features offer greater flexibility and room for experimentation. Importantly, it is already possible to retrieve data on the year of construction, floor area, number of floors and purpose of buildings via Estonian APIs. In this thesis, less granular building classes were also defined, and buildings were assigned their respective shares in each type, as described in subsection 3.3.1. It is worth noting that it was not possible to obtain metadata for all buildings as some of them were missing an EHR code. In general, data yet to be collected include average monthly temperature and number of occupants or information about the occupancy patterns.

6.2 Summary of the main contributions

One of the main results of this thesis was aimed at validating solar potential estimates, made in the previous work. To perform the validation, actual solar energy production data representing the total monthly generation for 17 buildings for three years (2020, 2021 and 2022) were obtained, cleaned and analyzed. Careful processing of the data, including removal of outliers, resulted in the calculation of monthly and annual solar power generation values that allowed us to see seasonal variations in generation. Finally, a careful comparison of solar energy potential estimates from the previous study with actual production data revealed discrepancies and showed the difficulty of validation, especially during winter months due to the lack of data. The validation step emphasized the need to improve the previous estimations to better match real production. In addition, it was found that the previously calculated azimuths did not correspond to the actual roof orientations, hence a new dataset with refined azimuths was created in this work.

To improve accuracy, filtering of rooftops as well as an approach for selecting more promising roof sections were proposed. Accurate roof orientation was achieved by correcting errors in roof azimuth calculations made in the previous work. Analysis of orientation changes after the recalculation of azimuths showed that these changes can increase or decrease production by 22% on average. Based on SmartEnCity buildings data, it was found that the actual roof area allocated to PV installations was about 60% on average, which is much lower than the previously assumed value. In addition, it was identified that north-facing roofs were not used for real PV installations as they are not promising for solar energy production. Aiming to overcome the limitations of previous estimates, the recalculation of potential Tartu's solar energy output was performed. The process started with building metadata mapping, including a suggestion of a convenient way to represent buildings' use purposes. In addition, data cleaning procedures were performed, leaving only more promising for solar panels installations rooftops. To optimize processing time of PVGIS API requests, a caching mechanism was implemented to eliminate redundant requests and reduce the computation time, improving efficiency and enabling future application of the methodology in larger cities.

Furthermore, it is important to note that the calculations were performed per one square meter simplifying further analysis and allowing for more flexibility in choosing the roof sizes for overall potential assessment. When validated on real production data from SmartEnCity buildings, the mean absolute percentage error (MAPE) was 11.82%, showing that the new estimations are closer to the actual production values.

In addition, the thesis reviewed methods for energy consumption prediction and analyzed the energy balance of Tartu based on available data. The investigation of regression models, as one of the most perspective approaches in the field, was conducted using the literature review. Taking into account the difficulty of accessing energy consumption data for model training due to privacy concerns, a methodology proposed Alessio Mastrucci *et. al* [18] was considered promising. This is because the approach utilized aggregated data at zip-code level, and in combination with additional features, it provided the foundation for energy consumption prediction. The energy balance analysis of Tartu revealed the potential for solar energy to cover between 48% and 97% of the total electricity consumption in the city, depending on the degree of roof coverage with solar panels. Finally, Tartu's solar energy production potential was performed. The analysis identified regions with high, medium, and low production clusters, visualizing color-coded buildings' footprints and revealing information on solar energy production capabilities across the city. These findings could assist policymakers and city government in formulating strategies for improving Tartu's sustainability.

6.3 Limitations and future work

The results of spatial analysis of Tartu's solar potential, outlined in the chapter 5, helped to gain valuable insights on solar capabilities of different areas across the city. Nevertheless, certain limitations have been discovered, which, in turn, opens up opportunities for refinement. It was found that in some areas, such as in Annelinn, buildings with flat roofs (based on a real map of Tartu) or buildings with a very similar type of construction were assigned to different production classes. The investigation of this issue revealed several underlying reasons. First of all, each building can have multiple roof sections, and based on the previous work calculations, these sections could vary in terms of tilt and azimuth. The differences could be attributed to various factors, such as roof design or certain functional structures on the roof that have been recognized as parts of the roof. Thus, if a flat-roof building has some sections with other slopes, its average production per m^2 will be lower and will be simply assigned with medium production class due to the upper quantile value taken as a threshold. As an example, the building on Kalda tee 4 in Tartu, has oblong pipes on the roof, distinguishing it from other buildings in the same area which are assumed to have similar design. Since the computations for identifying roof sections relied on 3D data in the previous work, the surface with the pipes was perceived as a sloped roof part. Furthermore, certain flat-roof buildings already hosted solar panels which were also treated as tilted roof sections. Consequently, the production potential of these buildings was in fact very close to the maximum potential, yet they did not surpass the threshold.

These limitations lead to several different suggestions for what can be improved in the future. The clustering of buildings based on production levels could be refined by selecting more optimal thresholds or considering the application of clusterization algorithms. This refinement would enhance the color-coded footprint map, offering a more precise spatial depiction of solar energy generation across Tartu. Moreover, an exploration into whether modifications to the data cleaning process or the roof section identification procedure could be made to ensure the exclusion or minimalization of structures that might mistakenly be perceived as roofs.

In a broader perspective, the spatial analysis itself could be improved to provide more comprehensive and detailed insights. Additional characteristics of city areas, such as number of people living there, could be taken into account when comparing the potential solar production of different districts. For instance, combining relevant data on population density with information on the number of buildings could provide insights into whether a district is more suburban in nature. Currently, the relationship between the type of city area and its potential contribution to the overall solar capacity remains uncertain due to the lack of such data. Nevertheless, this information could be valuable to better understand why some areas may produce excess energy, increasing their contribution to the city's overall potential, or, conversely, struggle to generate significant amounts of energy. By contrasting solar potential between zones characterized by small buildings and low population density with those featuring large buildings and high population density, a deeper understanding of sustainable urban design could be achieved, thus adding value to the inquiry of which city areas possess the greatest potential to contribute to the energy capacity.

Another direction for future work can be the prediction of energy consumption in Tartu. This thesis proposed models that can be used to achieve this goal and identified data collection requirements, laying the foundation for future research in this area. However, due to the different focus of this work and data unavailability, more precise methodology and its practical implementation were not developed. Thus, there exists a notable opportunity for further enhancement through the acquisition of relevant data and the subsequent development of a comprehensive energy usage prediction model. In addition, the electricity consumption values acquired in this thesis could be validated along with the resulting energy balance outlined in the section 4.3. It was assumed that the consumption information derived from the small amount of data available was correct. Nevertheless, it was not possible to verify the accuracy of the calculated energy balance values. Some doubts and skepticism about the results stem from the uncertainty that the balance could be as good as it was estimated to be. On the other hand, the balance was compiled solely on the basis of electricity consumption data, without taking into account the energy used for heating, for example, which leaves room for validity of the values obtained.

7 Conclusion

This thesis has made significant contributions towards improving the accuracy of Tartu's solar production estimations and understanding the solar energy potential of the city. The validation of previous solar potential estimates using real-world production data from the SmarEnCity project demonstrated the need for refining estimations. By rectifying errors in roof orientation and identifying roof sections with greater potential, the estimation methodology was enhanced. The implementation of a caching mechanism also improved computational efficiency by a factor of 5 compared to the previous approach, making the current methodology applicable to larger urban areas. Exploring energy consumption prediction methods, a promising approach utilizing aggregated data at the zip-code level, combined with additional features, was identified. The analysis of Tartu's energy balance revealed substantial potential for solar energy to cover a significant portion of the city's electricity consumption, ranging from 48% to 97% depending on roof coverage with solar panels. Spatial analysis visualised the recalculated Tartu's solar energy production potential, categorised into high, medium and low production capabilities. These findings can inform policy development and decision-making, facilitating the city's progress towards a more sustainable future. Overall, the validation of estimates, refined methodology, solar energy production analysis, and energy balance assessment collectively enhanced the understanding of Tartu's solar energy production potential.

The analysis and investigations conducted in this thesis revealed valuable insights regarding solar capabilities across the city. Nonetheless, certain limitations were identified, pointing to opportunities for refinement. One of the limitations pertains to the potential inflexibility of the threshold values utilized for classifying production. Other factor is related to the fact that rooftop constructions representing functional structures or existing solar panels were considered as roof sections in the previous work, resulting in a decreased overall production estimations for flat roofs. To tackle these issues, potential improvements include fine-tuning clustering through optimal thresholds or clustering algorithms and enhancing data cleaning and roof section identification procedures. Furthermore, the spatial analysis could be enhanced by considering additional city characteristics, like population density, to better understand the relationship between urban features and solar potential. Predicting energy consumption in Tartu presents another direction for future research, building upon the models and data collection requirements proposed in this thesis. By addressing these areas for improvement, a more comprehensive understanding of Tartu's solar capacity and its potential for sustainable urban energy utilization can be achieved.

References

[1] 9 Types of Buildings in Civil Engineering. https://dreamcivil.com/types-of-buildings/. Accessed: 2023.07.28.

[2] Pandas Library. https://pandas.pydata.org/. Accessed: 2023.07.22.

[3] Photovoltaic Geographical Information System (PVGIS). https://joint-research-centre.ec.europa.eu/photovoltaic-geographical-information-system-pvgis_en. Accessed: 2023.06.15.

[4] PVGIS API documentation. https://joint-research-centre.ec.europa.eu/photovoltaic-geographical-information-system-pvgis/getting-started-pvgis/api-non-interactive-service_en. Accessed: 2023.07.24.

[5] SmartEnCity – Towards Smart Zero CO2 Cities across Europe . https://smartencity.eu/. Accessed: 2023.07.22.

[6] Modeling and forecasting building energy consumption: A review of data-driven techniques. *Sustainable Cities and Society*, 48:101533, 2019.

[7] Eropean Environmet Agency. Urban sustainability: how can cities become sustainable? https://www.eea.europa.eu/themes/sustainability-transitions/urban-environment/urban-sustainability. Accessed: 2023.07.30.

[8] Eropean Environmet Agency. Urban sustainability in Europe - What is driving cities' environmental change? https://www.eea.europa.eu/publications/urban-sustainability-in-europe-what/. Accessed: 2023.07.30.

[9] Eropean Environmet Agency. Urban sustainability in Europe – opportunities for challenging times. https://www.eea.europa.eu/publications/urban-sustainability-in-europe/urban-sustainability-in-europe. Accessed: 2023.07.30.

[10] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[11] Center for sustainable systems University of Michigan. Photovoltaic Energy Factsheet. https://css.umich.edu/publications/factsheets/energy/photovoltaic-energy-factsheet. Accessed: 2023.07.30.

[12] S. Freitas, C. Catita, P. Redweik, and M.C. Brito. Modelling solar potential in the urban environment: State-of-the-art review. *Renewable and Sustainable Energy Reviews*, 41:915–931, 2015.

[13] Nelson Fumo and M.A. Rafe Biswas. Regression analysis for prediction of residential energy consumption. *Renewable and Sustainable Energy Reviews*, 47:332–343, 2015.

[14] Venkat Gudivada, Amy Apon, and Junhua Ding. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, 10(1):1–20, 2017.

[15] Tao Hong, Min Gui, Mesut E. Baran, and H. Lee Willis. Modeling and forecasting hourly electric load by multiple linear regression with interactions. In *IEEE PES General Meeting*, pages 1–8, 2010.

[16] Tae-Young Kim and Sung-Bae Cho. Predicting residential energy consumption using cnn-lstm neural networks. *Energy*, 182:72–81, 2019.

[17] Constantine E. Kontokosta and Christopher Tull. A data-driven predictive model of city-scale energy use in buildings. *Applied Energy*, 197:303–317, 2017.

[18] Alessio Mastrucci, Olivier Baume, Francesca Stazi, and Ulrich Leopold. Estimating energy savings for the residential building stock of an entire city: A gis-based statistical downscaling approach applied to rotterdam. *Energy and Buildings*, 75:358–367, 2014.

[19] Atika Qazi, Fayaz Hussain, Nasrudin ABD. Rahim, Glenn Hardaker, Daniyal Alghazzawi, Khaled Shaban, and Khalid Haruna. Towards sustainable energy: A systematic review of renewable energy sources, technologies, and public opinions. *IEEE Access*, 7:63837–63851, 2019.

[20] Bohdan Romashchenko. Mapping Solar Potential of Tartu. University of Tartu, 2022. Masters' thesis. Available at: `https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=75190`.

[21] Teolan Tomson. Discrete two-positional tracking of solar collectors. *Renewable Energy*, 33(3):400–405, 2008.

[22] Jian Qi Wang, Yu Du, and Jing Wang. Lstm based long-term energy consumption prediction with periodicity. *Energy*, 197:117197, 2020.

[23] Yixuan Wei, Xingxing Zhang, Yong Shi, Liang Xia, Song Pan, Jinshun Wu, Mengjie Han, and Xiaoyun Zhao. A review of data-driven approaches for prediction and

classification of building energy consumption. *Renewable and Sustainable Energy Reviews*, 82:1027–1047, 2018.

[24] Hai xiang Zhao and Frédéric Magoulès. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16(6):3586–3592, 2012.

Appendix

I. Code

The source code of this work along with all input and output data is publicly available at `https://github.com/amilisa/solar-potential-analysis`.

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Elizaveta Nikolaeva**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Analyzing the solar energy potential of Smart Cities,

   supervised by Pelle Jakovits.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Elizaveta Nikolaeva
11/08/2023