A GIS-based approach to estimate the geographic potential for rooftop solar PV: Developing a database at building-level for Denmark

Msc. Geoinformatics Master Thesis

Aalborg University (Copenhagen)

School of Architecture, Design and Planning (SADP)

Name:	Tinsley Wolfs
Student number:	20151525
Date:	09 – 06 – 2017
Length:	52 pages
Program:	MSc. Geoinformatics
Supervisor:	Carsten Kessler – Associate Professor
Semester:	SPLM MSC04

Abstract

The aim of this thesis is to develop a rooftop solar energy potential estimation for the country of Denmark - particularly a dataset containing individual buildings with attribute data on rooftop area in m^2, estimated to receive a range of solar irradiation in watt hours per square meter per day (Wh/m2/day). A technique often applied within potential estimation of solar energy at larger scales, involves a form of sampling. These samples are examined in detail and results at that scale are extrapolated. This approach is applied and based on the following assumption: Buildings of a comparable size and in a like context, will on average receive similar amounts of solar radiation in Wh/m2/day on similar proportions of rooftop area. However, a lack of time did not allow for a full calculation of the rooftop solar energy potential of the entire building stock. Therefore, the focus shifted to validating early results, to provide insight into and comment on the reliability of the often-used approach of sampling and extrapolation, in solar energy potential estimations.

Table of Contents

0. LIST OF TABLES, FIGURES AND GRAPHS

1. INTRODUCTION

Insight into the potential of solar energy across the stock of buildings at a large scale, can support renewable energy planning efforts at multiple administrative levels. The aim of this thesis is to develop a rooftop solar energy potential estimation for the country of Denmark - particularly a dataset containing individual buildings with attribute data on rooftop area in m^2, estimated to receive a range of solar irradiation in watt hours per square meter per day (Wh/m2/day).

The study limits itself to the solar energy potential on rooftops for several reasons. Roof space on buildings are commonly used to implement solar conversion technologies, such as solar heating and solar photovoltaics (PV). As it is elevated space and often with unobstructed vision of the sky. Furthermore, the implementation of these technologies on rooftops means that it is applied on existing land-use. Thus, it does not require space that forgoes other opportunities, while at the same time being near existing energy grids. Moreover, these buildings are owned by different actors that have varied means and incentives, impacting their ability and willingness to support the growth of renewable energy exploitation within the energy system, with their investment. On the other hand, there is the current lack of understanding of this potential. Previous large-scale studies conducted in the country of Denmark include Ahm and Vedde (2011) and Møller et al. (2012), which use very different approaches. And at the level of the individual buildings there exists a solar atlas called Sunmapper[1]. In addition, there is *Danmarks Meteorologiske Institut* (DMI), which provides country wide solar energy potentials based on meteorological data (DMI, 2012).

For the chosen study area, these existing large-scale estimates are rough and/or uncertain, while detailed potential estimations are provided for individual roofs only. Thus, there is a clear knowledge gap that this thesis aims to address. An additional aim is to provide for a degree of reproducibility. This is done through automation of the main data handling and modelling steps that are part of methodology with the use of scripting. It allows the process itself to be implemented more consistently and increase the transparency of the methodology and how it is applied.

Like in the cases of Møller et al. (2012) and Sunmapper, Geographic Information Systems (GIS) are used as a platform to support this thesis. It is highly suitable to support potential estimations with a strong spatial component and are heavily utilized in solar energy potential estimations. A technique often applied within potential estimation of solar energy at larger scales, involves a form of sampling. These samples are examined in detail and results at that scale are extrapolated. Due to practical limitations, this approach is applied and based on the following assumption:

- Buildings of a comparable size and in a like context, will on average receive similar amounts of solar radiation in Wh/m2/day on similar proportions of rooftop area.

However, meanwhile a method was devised and automated, a lack of time did not allow for a full calculation of the rooftop solar energy potential of the entire building stock. Therefore, the focus shifted

[1] Sunmapper project website: https://www.sunmapper.com

to validating early results, to provide insight into and comment on the reliability of the often-used approach of sampling and extrapolation, in solar energy potential estimations. The current structure of the report takes this shift of attention into account and has as guiding research question:

Is sampling a sound methodology in estimating the geographical potential of rooftop solar conversion technologies on a large scale?

The sub-questions that are to help answer this research question are:

- What is the state of GIS based potential estimation for solar conversion technologies?
- What variables are needed for an appropriate sampling strategy?

There are many ways in which potential estimations of solar energy conversion technologies, such as PV, are conducted. Accordingly, the first sub-question is important, because an assessment of the current state of GIS-based approaches, can help understand why the approaches differ and how this influences the result that interested parties, such as policy makers, rely on. The second sub-question is there to provide recommendations on what kind of data is needed to make reliable potential estimations considering the object under study, that of rooftops.

The structure of the rest of the thesis is as follows: First, is the methodology chapter, in which the existing literature is examined and the applied methodology is presented. Second, the results in which first the outcomes of the applied sampling strategy are discussed followed by a testing of the assumption. Third, a discussion chapter, where the methodology applied is evaluated along with the results it produces. Lastly, is the conclusion that returns to the research and sub-questions and to provide an answer as well as further recommendations for future studies.

2. Literature review and related work

This chapter examines the existing literature and related work in the study area of Denmark. First, it presents a hierarchical perspective frequently used in the practice of potential estimation of renewable energy resources. Second, the chapter assesses the developments in the potential estimation of solar energy, by describing the literature on solar radiation modelling. Third, an examination of the literature on potential estimation of rooftop solar photovoltaics (PV) is presented. This will give insight into the variety of methods applied in the equally varied contexts, which have influenced the methods used in this thesis. The next chapter will present this methodology.

2.1 Potential estimation of renewable energy resources

Hoogwijk (2004) presents a hierarchical perspective on potential estimation in her work on the global and regional potential of renewable energy resources[2]. At the base of this hierarchy is the theoretical limit of a primary resource. This is the energy of the sun reaching the surface of the earth for solar based conversion technologies.

The geographical potential of solar energy is dependent on the conditions at the location it radiates, such as: elevation, slope, aspect, material (reflectance) and context that may or may not cause shadowing effects. These factors cause spatial variation of solar radiation at the earth's surface. Accordingly, the geographical potential also relates to suitability of a location which is dependent on the applied conversion technology and imposed constraints. Usually, the aim is to estimate how much space in m^2 or km^2 is suitable within a study area, given an applied technology such as an average solar panel. Section 2.3, that examines the existing literature on potential estimation of rooftop solar photovoltaics, will show that this can be operationalized in many ways with significant differences for the estimation of the geographical potential (Ibid, 2004).

After the geographical potential comes the technical potential, which is the ability of a conversion technology to convert the primary energy resource into secondary energy, at a location judged suitable. For example, the potential kWh/m^2 a solar panel installation with a certain certified efficiency rate can produce over a period of a year. The economic potential can be derived from the technical potential and is based on the competitiveness with alternative energy resources and conversion technologies. This will determine what would be economically viable to implement.

The implementation potential is described as what can be implemented into the current energy system within a certain time frame and dependent on incentives and constraints, for example: governmental policies, aesthetical considerations and financial means (Ibid, 2004). This potential can be higher (or lower) than the economic potential, but never higher than the technical potential (Ibid, 2004). An example of when the implementation potential can be higher than the economic potential is when regardless

[2] Based on: Van Wijk, A. J. M., Coelingh, J. P. (1993). *Wind power potential in the OECD countries*. Utrecht University, Department of Science, Technology and Society, pp. 35.

of economic competitiveness, the choice is made to invest in sustainable energy conversion technologies like solar PV based on environmental considerations.

This thesis is concerned with quantifying the spatial variation of clear-sky global solar irradiation in Wh/m2/day on rooftop space across Denmark. Thus, it stops at the geographical potential. No specific conversion technology is used to determine potential in terms of total suitable area. The aim is to develop a base dataset that can serve a wide range of potential estimations of different solar conversion technologies, implemented on rooftops and in the context of Denmark.

2.2 Solar radiation modelling

Computational solar radiation models are based on empirical and meteorological research into the behavior of the sun and solar energy, and their interaction with the Earth and its atmosphere. This research examines the effects of atmospheric particles, slope, orientation, latitude, elevation, reflectance, obstructions that create shadows and the position of the sun in relation to the earth, to derive physically based and empirical formulas. Such efforts date back to the early 20th century and are reexamined and updated to this day (e.g. Ångstrøm, 1924, Shreve, 1924, Liu and Jordan, 1960, Robinson, 1966, Kondratyev, 1970, Nunez, 1980 and Scharmer and Grief, 2000). The developments in this literature are neither discussed in detail nor are the empirical formulas and how they have developed with time. Only the basic understanding of the factors that influence solar energy here on earth are provided.

In short, outside of the earth's atmosphere on a horizontal plane the energy of the sun is estimated to be is 1,367 W/m2 (Scharmer and Grief, 2000). Upon hitting the earth's atmosphere, it is reduced to 342 W/m2, due to the geometry of the Earth. This is further modified by the movement of the sun as seen from a position on the earth, which depends on the latitude, day of the year and time of day. Longitude influences the precise sunrise and sunset times in terms of Civil Time[3]. Solar radiation reaching the surface of the Earth is influenced by the length of the path through the atmosphere, cloud cover, and the particles that make up the atmosphere itself, which reflect, scatter and absorb part of this irradiance (Ibid, p. 34). At the surface of the Earth the location (latitude), slope, aspect, height and reflectance impact global irradiance at specific locations (Ibid, 2000).

Global solar radiation consists of three components: direct, diffuse and reflected radiation. Direct radiation is unaffected by objects obstructing direct vision of the sun, such as trees, buildings or cloud cover and in clear-sky conditions is the largest component of global radiation. Diffuse radiation is that which is scattered by the atmosphere and comes from all directions. When the sky is overcast the diffuse component makes up most of the global radiation measured at a location. Reflected radiation is reflected by different surfaces in the direction of the location where radiation is measured and is the smallest component of global radiation (Scharmer and Grief, 2000). Most of these factors are incorporated to a varying degree in solar radiation models.

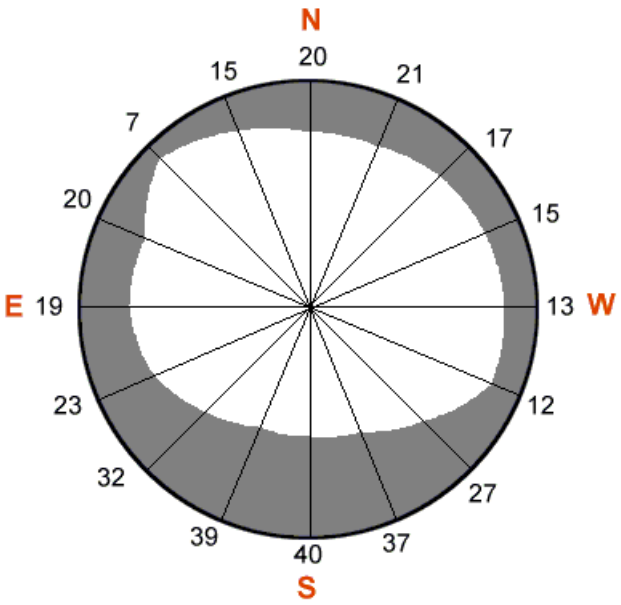
Early computational models make use of the complex formulas describing the above-mentioned interactions, to tabulate or graphically map solar radiation at locations of interest. Instead of

[3] In Solar Time or Local Apparent Time (LAT) sunrise and sunset times are independent of longitude and noon is when the sun is at its greatest elevation. In Civil Time or Local Mean Time (LMT) the sun rises and sets earlier by 4 minutes per degree latitude eastward in each time zone.

having to calculate them manually. Particularly of interest were those areas where meteorological data is sparse, like in mountainous terrain (e.g. Garnier and Ohmura, 1968 and 1970, Williams, et al 1972). However, the amount of points that could be calculated within a reasonable time frame were limited due to the capability of the hardware at the time. Thus, the results involved a degree of generalization across a study area.

In the late 1980s and 1990s three developments helped to increase the amount of detail of the modelling results. First, the introduction of Digital Surface Models (DSM) modelling complex terrain features that could be used as input data to model solar irradiation. A DSM is a grid of values indicating the elevation of a surface at each grid cell, with reference to a common base height, which is usually at sea level. The second is the increase of computational power to apply the complex formulas and third is the implementation of solar radiation models into GIS platforms allowing them to be used consistently and make the models more accessible (Hetrick et al, 1993, Dubaya and Rich, 1995, Kumar et al. 1997).

Two such models implemented in well-known GIS platforms are rsun for the open source platforms of GRASS GIS and QGIS, and Solar Analyst provided in the Spatial Analyst extension of the commercial GIS platform ArcGIS. There are other solar radiation models available, such as DAYSIM[4] (Reinhart, 2001, used in Jakubiec and Reinhart, 2013). These are not examined in detail here, but are other options that can be explored for solar radiation modelling. Solar Analyst has gone through several iterations, with first the SOLARFLUX model (Hetrick et al. 1993), then by the model developed by Kumar et al (1997) and in early 2000 by its current implementation developed by Fu and Rich (1999, 2002). r.sun was first introduced in 2004 by Hofierka and Suri (2004, 2007) and implemented in the GRASS GIS platform (Neteler and Mitasova, 2008). A new iteration is presented in Hofierka and Zlocha (2012) called v.sun, which can process 3D vector data instead of a DSM.



1E) The resulting viewshed for a location represents which sky directions are visible and which are obscured. Numbers represent the calculated horizon angles.

Figure 1 Visualization of the viewshed methodology (Fu and Rich, 1999, p. 6)

[4] DAYSIM: http://daysim.ning.com/

Both the r.sun and Solar Analyst use a DSM as their main input, but the main difference is how they account for topographical features that create shadows. The Solar Analyst model calculates a 360 degree viewshed on each grid cell and the procedure is described by Fu and Rich (1999 and 2002), as similar to lying down and using a fish-eye lens to capture the immediate surroundings. The model provides a parameter where the user can specify how many directions the surroundings are sampled, from which it reconstructs the horizon. The more directions it can sample the surrounding from a single cell, the horizon can be constructed with more detail. The result would be used to modify the amount of solar irradiation on the calculated location in those directions where the horizon blocks, fully or partially, the direct radiation from the sun. Figure 1, from their paper, visualizes this horizon. The dark area is what obstructs direct solar radiation, while the white area would indicate unobstructed vision of the sky and the sun.

The Solar Analyst model calculates both direct and diffuse radiation, not the reflected component. Because Fu and Rich argue that this is only a small component of the global radiation and modelling the reflectance of a surface, which depends on the properties of the latter, is a complex interaction to account for (1999). Direct solar radiation is calculated on the basis of a so-called sun map. It maps and tracks the position of the sun across the sky throughout the day and year based on the latitude of the study area and specified time range. The time range, defined in day(s) of the year and hourly time step define from how many positions it will calculate direct radiation, which are represented by sectors.

Figure 2 is an example of a sun map, which each sector represented by a distinct color. It visualizes the directions from which direct solar radiation will be calculated between June 21st and December 21st. The previous image, showing the horizon is places on top of this sun map and acts as a modifier of direct irradiation, reducing the amount of irradiation coming from those specific directions.

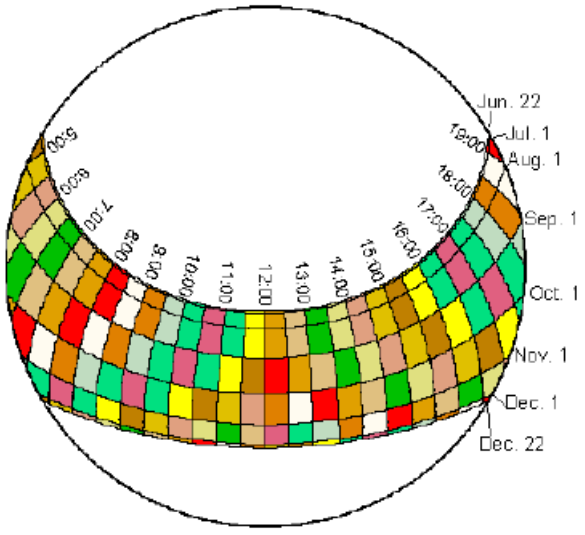


Fig. 2. Annual sunmaps for 39° N latitude using 0.5 hour intervals through the day and month intervals through the year.

Figure 2 Example of a sun map (Fu and Rich, 1999, p. 7)

Diffuse irradiation comes from all parts of the sky, obstructed and unobstructed, due to the scattering of radiation by the atmosphere (and clouds). To calculate the diffuse component the model creates a so-called sky map (figure 3). It also consists of multiple sectors, representing different zenith and azimuth angles for which it calculates the diffuse component of solar radiation that it contributes to the radiation at each cell. When more sectors are specified, the calculation will take longer to complete, but the results will be more precise. This sky map is also modified by the horizon map, by overlaying the former with the latter, which reduces the diffuse radiation coming from those directions.

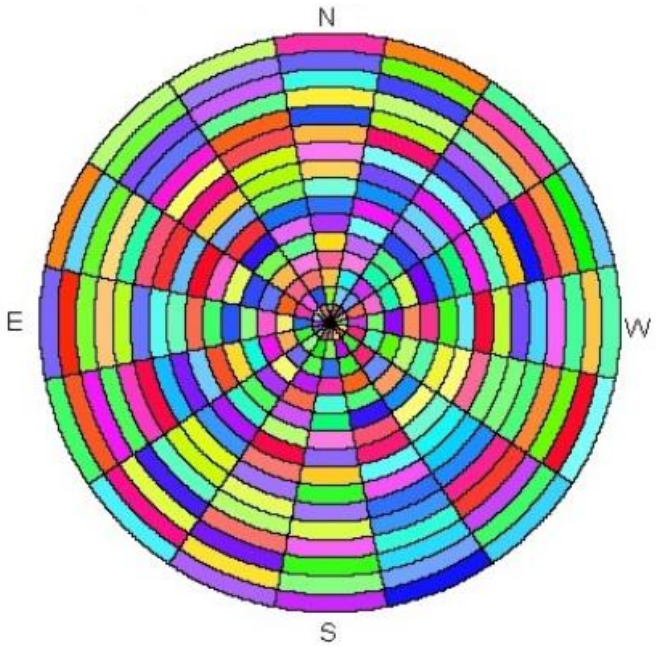


Figure 3 Example of a sky map (Fu and Rich, 1999, p. 8)

The effects of other atmospheric attenuators of solar radiation can be set through two additional parameters. One is transmittivity and the other the diffuse proportion of global radiation. With the latter, a static modifier is applied to the global radiation reducing direct radiation component of global irradiation or increasing it (to mimic cloud cover or real-sky effects). Transmissivity is the ratio between energy reaching the earth's surface and what is received at the upper atmosphere, the lower this value the more direct radiation is scattered or absorbed by the atmosphere and vice versa. Transmissivity and the diffuse radiation component have an inverse relationship, if the former increases the latter decreases.

The rsun model of Hofierka and Suri (2004) base its formulas on the European Solar Radiation Atlas (Scharmer and Grief, 2000). They argued at the time that existing implementations, like Solar Analyst, are build either on a simplified representation of the fundamental physics, or do not provide flexibility when it comes to the input data that reflect cloud cover and other atmospheric attenuators. For example, instead of setting a constant value mimicking the effects of atmospheric transmissivity and cloud cover, for the entire study area, it is possible to input raster data that can spatially vary these effects. Moreover, instead of using a fixed latitude, this is assessed per grid cell, which means that across larger areas where latitudinal differences have an impact, the latter's effect is considered.

Furthermore, it offers the ability to calculate the reflected component of global irradiation. But its ability to do so is limited to the DSM that is used, which captures height values and vertical surfaces are not as clearly modelled, since the height values recorded mainly parallel to the ground. But with its inclusion it is possible to set either a constant reflectance value (albedo) or input a raster in which this is spatially varied. The Solar Analyst does not offer this option, but it will allow users of rsun to try and approximate the reflected component, which in highly urbanized areas is likely more prevalent.

Similar to the Solar Analyst model, the user inputs a time frame in days, and a time step in hours, or a fraction of an hour. Based on the user preferences it can output a surface with direct, diffuse, reflected or global radiation values, or all of them, for each time step, or aggregated across the set time frame. Like the Solar analyst model, it also calculates a horizon that models the surroundings which may or may not obstruct a clear view of the sky and/or sun, but does so in a different manner. The method is as follows: for each location on a DSM it compares the height values of the adjacent pixels in a set amount of directions. If the adjacent pixel is higher than the starting pixels it increases the angle, until that angle is higher than any adjacent cell in that direction, or if it reaches the border or a study area. If this is done for a large study area, it considers the curvature of the Earth. The output is a surface with horizon angles in degrees separately for each requested direction. The more directions the horizon is sampled, the horizon will be constructed with more detail but this will increase the processing time. Figure 4 below shows an example of modelled solar radiation values on top of a DSM. On the left is a satellite image of the same area.

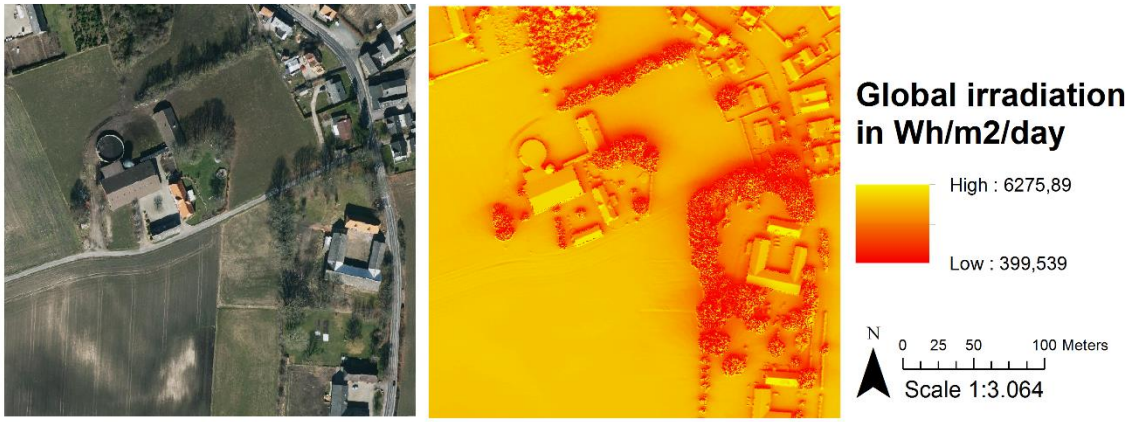


Figure 4 Example output of the rsun model

2.3 Estimating the geographical potential of rooftop solar energy

This section focusses on research done on the potential of rooftop solar PV, as it will be illustrative of how the geographical potential at different scales is conceptualized and derived. The methodology applied depends on data availability, data quality, size and context of the study area and practical limitations such as: time, storage capacity and hardware capabilities. This section describes several studies in detail to show how varied the different methodologies currently are. Studies that are similar to those discussed in detail are also mentioned, but some will contain aspects similar to others.

Two studies that indicate a divergence in the way a geographical potential is calculated are the study of Lehmann and Peter (2003) and Suri et al. (2007). The former is a statistically based approach,

while the latter makes use of a land use database called CORINE (CLC90) and GIS. Lehman and Peter (2003) correlate usable area and population density to assess the roof and façade potential for solar use across the Europe Union (EU15). They base usable area on a study performed by Gernhard et al. (1992)[5] into roof and façade areas of residential and non-residential buildings in Northrhine-Westphalia, Germany. They assume this data to be valid and extrapolate the results to the rest of the EU15 based on the assumption that "[…] building and living structures in western European countries are primary dependent from local population's densities and that they are similar if population density is […]" (p. 1).

The statistical approach taken is to fit several curves between population density in capita/km^2 on the one hand and roof or facade area in m^2/capita on the other, for both non-residential and residential buildings from low to high density. From there they estimate based on different densities a range between 4-9 m^2/capita of available roof area and 3-5 m^2/capita of available façade area on non-residential buildings. On residential buildings, they estimate a range 4.5-7 m2/capita of available roof area and 5-6 m2/capita of façade area. Applied to the area of the EU15 they estimate a total available area for solar use of 7,000 km^2 of which 4,600 km^2 is on roofs and 2,400 km^2 is on facades. In the process, they apply a factor of 0.9 for roof areas and 0.66 for facades to mimic losses due to unusable areas and shadowing effects. For Denmark, they estimate the total available roof and façade areas to be 65 km^2 (12.7 m^2/cap.) and 35km^2 (6.8 m^2/cap.) for low and high-density areas respectively. There are a several studies that use the same relationship between population density and available area to extrapolate results, for example: Izquierdo, et al. (2008), Bergamasco and Asinari (2011a), Wiginton, et al. (2010), Khan and Arsalan (2016).

On the other hand, Suri et al. (2007)[6] use the validated PVGIS database (at 1km^2 resolution) to estimate the potential of solar electricity generation for the European Union and several adjacent countries (EU25 and Bulgaria, Croatia, FYROM, Romania and Turkey). They considered the existing building infrastructure in urban residential areas represented by class 11 (urban fabric) of the CORINE Land Cover (CLC1990) database which has a 100m^2 resolution, and excluded industrial and commercial areas as well as transport infrastructure and parks. They aggregate the results to the administrative boundaries of Eurostat NUTS level 3 for each country to assess regional differences[7].

Suri et al. (2007) does not define geographical potential as total suitable area, but use the area defined by CORINE and compares potential solar electricity production of both horizontally-mounted PV modules and at optimal inclination, on these areas across Europe. In addition, they estimate the area necessary to meet the national electricity consumption through solar PV (based on IEA, 2004). For the latter estimation, they use an average 1kWp system consisting of PV modules with an area of 9.5m^2. For Denmark, they estimate that 0.91% of the surface of the country would be required, to cover the electricity demand, ignoring cross-border electricity trade. A similar study is that of Nguyen and Pearce (2010), Redwiek et al. (2013) and Sun et al. (2013).

A combination of these two approaches is provided by Izquierdo et al. (2008). Izquierdo et al. (2008), estimate the technical potential of rooftop PV for the country of Spain. They recognize that size of the study area is an important aspect requiring different techniques at different scales. Therefore, they turn to the use of homogenous averaged data. For the theoretical potential, their input data is a DEM with a resolution of 90m^2 and meteorological data from 110 stations spread across the Iberian Peninsula. These are

[5] Gernhard, D., Mohr, M., Skiba, M., Unger, H. (1992). *Theoretisches und technisches Potential von Solarthermie, Photovoltaik, Biomasse und Wind in Nordrhein-Westfalen*. Ruhr-Universität, RUB-E-I-35

[6] Huld, Müller and Gamardella (2012) is a follow up publication containing updates to the PVGIS database.

[7] For a comparison of the different administrative levels represented by the NUTS classification (Nomenclature of territorial units for statistics), see: http://ec.europa.eu/eurostat/web/nuts/overview

used to calculate the global horizontal irradiation in kWh/m^2/month across the study area, which is used to calculate the technical potential on available rooftop space.

To derive the geographical potential a stratified statistical sampling approach is applied to define a so called 'representative building typology' (RTB). The RTB describes average building characteristics within each municipality and acts as homogenous averaged data to operationalize the estimate at country scale. They mention that the characteristics of an RTB include: typical building height, roof structure or building compactness, but differentiate between RTBs based on population and building density as they assume that these determine the typologies (similar to Lehmann and Peter (2003). Based on both population and building density[8], they classify each municipality from low density to high density on these two axes. Afterwards they take samples from each group of classified municipalities based on a confidence level, in which they select relevant built-up surfaces from vector data based on several classes of CORINE (CLC 2000 at 200m^2 resolution) and measure the total built-up area.

To define available roof area, they apply a range of coefficients to the measured relevant built-up area, consisting of a: void fraction coefficient (defined as "[…] voids and recesses in buildings" (p. 933), shadowing coefficient and a facility coefficient (mimicking existing roof uses). The shadowing and facility coefficient are determined based on human inspection of satellite imagery in the sampling areas, and the void fraction coefficient based on the same process as deriving the built-up area. The results are extrapolated to other municipalities within the same class, with a degree of variation based on the size of the municipality itself. With the inclusion of an a priori error margin they estimate that there is a total built-up area of 2,884 ± 231km^2 and available roof area of 571 ± 183km^2, which per capita translates to an average of 14.0 ± 4.5 m^2/capita. Comparatively, Lehmann and Peter (2003) estimated a total area of 528 km^2 and measured 13.5 m^2/capita of available roof area in Spain.

Other examples which also define coefficients in various ways to determine usable roof area as the geographical potential are: Scartezzini et al. (2002), Montavon, et al. (2004), Gosh and Vale (2006), Pillai and Banerjee (2007), Wiginton, et al. (2010), Ordonez, et al. (2010), Bergamasco and Asinari (2011a,b) and Khan and Arsalan (2016). It is important to note that in many cases the theoretical potential does not factor into the geographical potential in large scale studies. The former is used to derive the technical potential but its spatial variation has no bearing on defining the geographical potential, only the mentioned coefficients (in particular the shadowing coefficient) and the sampling and extrapolation strategy.

The study from Hofierka and Kanuk (2009) in which they estimate PV potential at the scale of a part of a city in Slovakia (Bardejov) with a 3D model, is comparatively illustrative for the difference in technique that is applied at a smaller scale[9]. The obvious advantage is the ability to capture local conditions which Hofierka and Kanuk (2009) model with the use of a high-resolution DEM and vectorized orthophotos which both are at a resolution of 1m^2. At this accuracy shadowing effects can be modelled with high accuracy within urban areas compared to that of the PVGIS database which models these at a resolution of 90m^2, for Europe and or visual inspection of applied by Izquierdo, et al. (2008), for Spain. In addition, with the use of a vectorized orthophotos building objects are clearly defined in comparison to using CORINE, which represents outlines of built-up area not individual buildings (used in Suri, et al., 2007).

Hofierka and Kanuk (2009) define theoretical potential as the spatial and temporal variation of solar irradiation on both roof areas, and in between, in the modelled city and use the same r.sun model as

[8] The number of inhabitants and building count are divided by the surface area in km^2 of each municipality.

[9] In the studies mentioned that make use of a DEM or DSM are essentially 2.5D models not 3D models. None of such studies model gaps underneath elevated terrain which is a characteristic of a full 3D model.

Suri et al. (2007). The geographical potential is limited to the building footprints and is divided between urban zones they specify as: residential houses, blocks of flats, industrial areas and other facilities. For each building in the study area they add a range of attributes such as: building function, roof type, inclination, orientation and height, as well as its footprint area, roof area and roof area available for PV. They measured these attributes using a laser distance device. A minimum available roof area for PV of $10m^2$ is used as a threshold to include a building into the potential estimation. In total, they estimate that 34.9% of the total building footprint of residential housing is available for PV systems, while this is 75% for blocks of flats, 60.9% of industrial areas and 66.67% for the other facilities.

With the use of three classes from the CORINE (CORINE 2000) to locate urbanized areas across Slovakia[10], they extrapolate the results to the rest of Slovakia. The extrapolation assumes that the mentioned zones are found in all major cities due to "[…] the uniform architectural design of buildings built during the socialist era in the second half of the 20th century" (p. 2208). In total, they mention that 8.1% of the urban areas in Slovakia can directly be used to produce electricity[11].

Examples in between the aforementioned scales are the regional studies of Bergamasco and Asinari (2011 a-b). These are also indicative of the impact of applying different techniques in a similar context. Bergamasco and Asinari estimated the potential of solar PV in the Piedmont Region (2011a) and the city of Turin (2011b) in northern Italy. In their first study data on the theoretical potential is based on the PVGIS database in form of a raster with year sum values of global irradiation on optimally inclined surfaces (kWh/m2) and another with values of optimal inclination angles of equator facing PV modules. They took a mean global radiation value from the pixels that fall within each municipality.

To assess the geographical potential, they use a multitude of vector maps at a scale of 1:50,000, from which they extract the entities that represent residential and industrial buildings, used to measure the total roof area per municipality. Afterwards, the available roof area is derived in similar fashion as presented in Izquierdo et al. (2008) through the application of a range of coefficients. They similarly define a RBT as they do not know the roof characteristics and no information is available to validate their results. Instead of it being based on population and building density, which according to Izquierdo et al. (2008) defines the building typology within an administrative unit, they base it on the architectural characteristics of an average residential and industrial building directly. In the table below the coefficients are presented for residential and industrial buildings. Except for the coefficient 'shadowing effects of neighboring buildings' which they took from Izquierdo et al. (2008), Bergamasco and Asinari based the others on visual inspection of Google Earth imagery.

[10] They used the classes: 111 – Continuous urban fabric, 112 – Discontinuous urban fabric and 121 – Industrial or commercial units. (p. 2210)

[11] A simplified approach to small scale technical potential estimations on rooftops is provided by Lee, Hong, Koo, Jeong and Kim (2016). They use a hillshade analysis to measure shadowing effects on simplified building polygons. They additionally a minimum amount of available free roof area to place a standard solar PV panel, to determine the total geographical potential.

Table 1 Coefficients modifying residential and industrial buildings rooftop area (Bergamasco and Asinari, 2011a)

Coefficient	*Residential buildings*	*Industrial buildings*
Double pitched roof	0.5	0.75
Existing roof uses	0.7	0.9
Existing solar thermal installations	0.9	1.0
Shadowing effects of PV arrays on itself	0.45	0.45
Shadowing effects of neighboring buildings	0.46	1.0
Roof inclination	20%	30%

In their follow up study on the city of Turin (2011b), they present an updated methodology whereby they exchange the visual inspection of Google Earth imagery to determine the coefficients, with a systematical analysis of geo-referenced ortho-imagery retrieved between 2002 and 2004, to evaluate shadowing, available roof surface (based on brightness), existing roof uses and azimuthal angle of each individual building. The inputs are 124 ortho-images each of roughly 2GB in size from which they extract an image of each industrial and residential building using the vector data of their previous study (2011a). This is to be analyzed with the use of a high-performance computer cluster.[12] They mention that the timeliness of the imagery may not accurately reflect reality and thus can introduce errors in the assessment. They quantify the accuracy of the applied image analysis technique to be 90%. Their new coefficients, for which they used a slightly different naming scheme can be found in the table below.

Table 2 Updated coefficients modifying residential and industrial buildings rooftop area (Bergamasco and Asinari, 2011b)

Coefficient	*Residential buildings*	*Industrial buildings*
Shadow	16.3%	12.7%
Suitable (not bright)	38.0%	39.5%
Suitable (bright)	43.1%	46.0%
Features (existing roof uses)	0.9%	0.8%
Mean azimuthal angle	9°	7°

Interesting is the decrease in the coefficient on shadows. By applying a more systematical approach Bergamasco and Asinari measured a significant reduction of the impact this coefficient has. It illustrates that in a similar context, the results can differ markedly based on using methodology that derive homogenous averaged data, or methodology that can assess the full heterogeneity of the study area. This difference is also reflected in their potential estimation, which increased by 41% compared to the results of their previous study. Similar studies to that of Bergamasco and Asinari (2011b) is Kabir et al. (2010), Latif et al. (2012) and Rodriguez et al. (2017).

Jakubiec and Reinhart (2013) present another combination of techniques at the scale of the city of Cambridge, Massachusetts, USA. They do not explicitly use the hierarchical perspective presented in

[12] From their paper: ”It is a Transtec HPC cluster of 72 total virtual cores, 144GB total RAM and 5.5 TB total hard disc capacity” (p. 2744).

Hoogwijk (2004). But to provide a point of comparison they combine the theoretical, technical and economic potential as a direct influence on what is constituted as suitable in terms of geographical potential. Like Hofierka and Kanuk (2009) they construct a detailed 3D model of the 17,000 buildings in the study area out of LiDAR data at a resolution of 1.25m^2 and exclude roof area that has a slope of more than 60 degrees. They quantify the accuracy of this model to be below 1m^2 RMSE. For the theoretical potential, they also model solar irradiation but instead of using r.sun they make use of the Daysim simulation software. They calculate annual irradiation values at a resolution of 1.5m^2 on roof surfaces with the direction and inclination based on the 3D modelled surfaces which they validate against meteorological data.

The technical potential is derived using an average solar panel rated at 185W/m^2 of which they validate the results against a sample of two existing solar PV installations in the study area. The economic potential is based the ability to break even on the investment after a 10-year period with a 10% discount rate per year with fixed local electricity costs for the duration. A combination of the three, and excluding any other financial incentives, indicated that at those locations which can generate 1,244.9 kWh/m^2/year or more are deemed suitable and would be part of the geographical potential. With the inclusion of existing financial incentives, such as a multitude of national and state tax rebate programs, locations generating 121 kWh/m^2/year or more would already be deemed suitable. This considerably influences the suitable area estimating the geographical potential. The results of their study were built into an online accessible database[13] which is one of many constructed for other cities across the USA[14]. Similar studies include: McIntyre (2012), Norbert (2012), Li, et al (2016) and Brito, et al (2017).

An example when considerations as time, storage, computational power, data availability and scale do not form barriers is Project Sunroof developed by Google (2017)[15]. The methodology applied to develop this database also acts as an example of another technique more recently applied, which is machine learning. The database focusses on providing a technical potential for each address using an average panel rated at 250W with a size of 1.650m by 0.992m. Additionally provided, is a simple savings estimator and links to solar panel installation companies that service the locations queried. Furthermore, the results can be viewed at different administrative levels such as a city or state. At the time of writing the database consists of about 60 million buildings situated in 50 states and Washington DC (Google, 2017).

Similarly to Jakubiec and Reinhart (2013) and Hofierka and Kanuk (2009), they develop a 3D model of each building and its surroundings up to 150 meters at a high resolution (the exact resolution is not mentioned, but likely to be below 1m^2). Google's own rough outlines provided in Google Maps represent the building footprints, which they further process using a Deep Learning neural net (details are not provided) in combination with aerial imagery to identify green objects, such as trees overhanging parts of a roof, and surfaces at ground height. The latter two are removed to create a clear outline of the actual roof area and existing roof uses. Like to Bergamasco and Asinari (2011b), the timeliness of the aerial imagery can introduce a degree of error as well as that the algorithm applied may falsely identify other structures as buildings.

The theoretical potential consists of a calculation of direct normal irradiance and diffuse horizontal irradiance values, validated with meteorological data provided by NREL, and take into account shading effects of neighboring objects and features of the roof itself. The validation is, however dependent on the weather station data of which spacing may not be optional in relation to the modelled location. With

[13] Link to online database: https://www.mapdwell.com/en/solar/cambridge.
[14] See table 1 on page 128 of Jakubiec and Reinhart (2013). Although, these databases are developed with varying methodologies.
[15] Google Project Sunroof (2017): https://www.google.com/get/sunroof#p=0.

the use of the RANSAC algorithm they identify contiguous segments on each rooftop, as well as their orientation and slope. With that information, they use a greedy algorithm that maximizes the total energy received and size of the solar panel array on the identified contiguous roof segment(s). At minimum, a roof must house a capacity of 2KW and a single roof segment needs to have enough space to place at least four panels, for the roof to be included in the database as suitable. Another recent study that makes use of a machine learning approach is Assouline et al. (2017).

2.4 Related work in the area of study

In the introduction, several related works in the same area of study are presented, which are now described in more detail. First is Ahm and Vedde (2011), they discuss the potential of large scale solar PV plants over 200kW in Denmark and do not cover implementations on private housing. Instead, they focus on:

- flat roofs of large commercial buildings,
- large scale infrastructure such as noise barriers and
- ground mounted PV installations

They estimate the potential of these three options in terms of space by examining land-use. Concerning buildings, the authors say that these have a footprint corresponding to 600km^2 and constraints, such as orientation and shadowing, will reduce the area available to implement PV to a quarter. Of that quarter, they project that 50km^2 consist of large scale flat roofs on commercial buildings that can be used for large scale plants. But they mention that the actual potential in terms of km^2 of suitable large-scale roofs in Denmark is not known.

Møller et al. (2012), present a methodology behind a solar atlas of 2.9 million buildings, combined with the national building registry, in Denmark. Inspired by existing solar atlases such as: German SunArea[16] and Bristol Solar Map[17], they argue that such an atlas can help in "… assessing the use of PV at larger scales by means of quantitative data on potentials as well as their associated costs." (Møller et al., 2012, p. 1). Like Ahm and Vedde (2011), they estimate the total footprint area to be roughly 600km^2 and mention that if a third was available, in theory the technical potential is in the 20GWs.

Though, instead of using land-use, they make use of a high-resolution digital elevation model (DEM) at 1.6m by 1.6m resolution. A digital elevation model is a grid of cells that specify a height value in relation to a common base height, which is usually at sea level. With the use of this DEM, they intended to calculate and map solar radiation on all roofs, but ran into technical and time related issues where the full set of computations could not be conducted for all municipalities in Denmark. Instead, they present the results of Aalborg and the greater Copenhagen area in their work. By using a DEM, it allowed them to take such variables as location, orientation and inclination of individual roofs, and shadowing caused by neighboring objects into account, which land-use cannot.

[16] German SunArea: http://www.sun-area.net
[17] Bristol Solar Map: https://www.bristol.gov.uk/housing/solar-panels

Since 2015, a solar atlas for rooftop solar PV potential exists for the country of Denmark called Sunmapper[18], developed by the Danish Technical University (DTU) in collaboration with several municipalities and Climate KIC. On their website users can type in an address and the server calculates the solar PV potential on a DEM of 40 by 40 centimeters of the roof at that address. Similar to Møller et al. (2012) it considers neighboring objects but adds local weather data provided by DMI. However, no option is available to calculate the potential of an entire region or municipality. Its aim is to provide incentives to private actors to invest in solar PV by giving insight into the potential of their property.

DMI provides maps at country scale indicating global solar radiation in yearly megajoule per square meter total (Mj/m2) but at a low resolution (DMI, 2012) or measurements of several weather stations spread across the country. These maps are large with a low resolution, which means that the impact of the spatial structure of developed areas and individual roof characteristics are not accounted for. However, this data can be used as input data in the solar radiation modelling process to determine real-sky global solar radiation or to validate modelled results at a large scale.

Suri et al. (2005) present a solar radiation database called the Photovoltaic Geographical Information System (PVGIS) for Europe that includes maps of Denmark. The database is in the form of multiple rasters developed with the use of the r.sun solar radiation model. With the addition of several tools accessible online, allow the user to estimate PV electricity production at point locations[19], thus providing a theoretical, geographical and technical potential. This database covers most of Europe with a size of 5000 km by 4500 km (p. 56). Their input data is a digital elevation model (DEM) with a resolution of 1 km^2, along with data from 566 meteorological stations on monthly averages of daily sums of global and diffuse irradiation during 1981-1990 and linke turbidity values from 611 locations.

The database consists of a raster with daily sums of global irradiation on a horizontal surface in $Wh/m^2/day$, a raster with daily sums of global irradiation for surfaces inclined at 15, 25, 40 and 90 degrees in $Wh/m^2/day$ and one with monthly and one with yearly average values of the daily sums of global irradiation for surfaces at an optimal surface inclination in $Wh/m^2/day$. By using a DEM they are able to model the effects of shadowing caused by larger terrain features, however the resolution is coarse at $1km^2$, thus it loses accuracy at larger scales.

The PVGIS project is still being developed and is currently in its fifth iteration[20]. Over time they have extended the coverage to Africa, Asia and now North, Central and South America. In addition, they recalculated solar radiation and PV energy yield estimations with meteorological measurements from satellite data from CMSAF, SARAH and data provided by the National Renewable Energy Laboratory (NREL). In addition, they increased the temporal resolution to hourly, included the cooling effects of PV panels by wind and increased the spatial resolution to $90m^2$ (Huld, Müller, Gambardella, 2012, Huld, et al 2017). This means that it much less coarse, however it would not be able to take account of objects smaller than that resolution which means that it is still to coarse for the aim of this thesis.

[18] Sunmapper: https://www.sunmapper.com
[19] This application can be accessed on the following website: http://re.jrc.ec.europa.eu/pvgis.html
[20] http://re.jrc.ec.europa.eu/PVGIS5-beta.html

2.5 Concluding the literature review

To reiterate, methodologies for potential estimation of rooftop solar PV are influenced by data availability, data quality, size and context of the study area, and practical limitations such as: time, storage capacity and hardware. The examination of the literature on solar PV potential estimation has shown that a combination of these factors influences the methodologies applied. When detailed building outlines are not available they must be constructed for the study area, such as seen in Bergamasco and Asinari (2011a). Or they are constructed based on a sample for practical reasons and extrapolated from (e.g. Izquierdo et al., 2008, Wiginton et al., 2010). Another approach is to use data that represents built-up area such as land-use (e.g. Suri, et al., 2007), which is often the only comparative data available at very large scales that encompass multiple countries.

When it comes to data quality the use of building outlines will be more precise compared to land-use when estimating the total built-up area from which available rooftop space is to be estimated. However, when detailed ortho-imagery is available that can be analyzed by image classification techniques (e.g. Bergamasco and Asinari, 2011b), or there is the possibility to develop a 3D model that can be analyzed (e.g. Jakubiec and Reinhart, 2013); rooftops and available space are defined with much higher accuracy.

The scale currently seems to dictate the choice of methodology. Smaller scale studies resort to detailed 3D models of all the relevant objects with moderate to very high accuracy (e.g. Hong, et al., 2016 and Jankubiec and Reinhart (2013). At larger scales the heterogeneity of the building stock is abstracted to a homogenous average (e.g. Izquierdo, et al., 2008, Bergamasco and Asinari, 2011a). The definition of this homogenous average that represents the building stock will have the most influence on the eventual potential estimation in many studies, that deal with a larger scale and need to consider practical limitations.

To define a homogenous average, studies try to derive coefficients that are assumed representative of roof structure and contextual influences. These are used to modify the measured total built-up or building footprint area to arrive at usable or suitable rooftop space for PV applications. But the methods to define these coefficients differ markedly. From purely statistical exercises and visual inspection of satellite imagery, to advanced image analysis and 3D modelling and more recently the introduction of machine learning. They result in strong differences on what is identified that impacts suitable rooftop area and how much.

Such strong differences are likely an indication of the impact context has in which a study is performed. A closer examination of local conditions through any methodology, where possible, shows that there is a high degree of heterogeneity in the urban morphology and architectural characteristics between researched areas. This seems explanatory of the broad range of coefficients that can be found in the literature. But this may point at another relevance of scale. Beside its impact on methodology, when are practical limitations are to be considered, it is of importance in the process of defining a homogenous, representative average of the heterogeneous building stock. If the building stock can be assumed heterogeneous, then it is likely that as the scale of the study area increases heterogeneity increases (which further increases if time is considered).

A case in point is the two studies performed by Bergamasco and Asinari (2011a-b). The applied coefficients underestimated the potential in the city of Turin by 41%, compared to a detailed examination in their follow-up study. Thus, if the size of the study area introduces practical limitations, which requires the building stock to be sampled to define a homogenous average as a set of coefficients, then the scale at which

this average is defined needs to be considered and tested to minimize estimation errors. The scale at which the homogeneous average is defined of the Bergamasco and Asinari's first study may have been too large and thus not representative of local conditions. Perhaps an approach where multiple averages representative of parts of the study area would have performed better, as applied in Izquierdo, et al. (2008).

Currently, the main practice to stratify the heterogeneous building stock is the relationship between population density and (available) roof area presented in Lehmann and Peter (2003). The results of Lehmann and Peter (2003) and Izquierdo et al. (2008) are not far removed from each other, which seems to indicate that it works. However, there is a lack of validation if stratifying based on this relationship, does account for the local heterogeneity of rooftop characteristics and urban morphology (e.g. height, slope and roof orientation, existing roof uses and contextual influences such as neighboring objects).

The next section of this chapter presents the methodology applied in this thesis. Like Izquierdo et al. (2008) the aim is to construct multiple representative homogeneous averages, due to practical limitations and scale of the study. The ability of the stratification to produce representative averages is additionally tested to see if extrapolation can be done reliably to provide insight into the suitability of using density to both stratify and upscale potential estimations.

3. Methodology

This section will describe the methodology that is to be applied in this thesis to estimate the geographical potential of solar energy, in the context and at the scale of the country of Denmark. It is assumed that the building stock is heterogeneous and the aim of the methodology presented here is to take this into account with the available data. First, several practical limitations must be mentioned that influenced the methods that could be applied within a reasonable time span. These limitations include time, storage capacity and hardware capabilities.

This thesis was conducted over a four and half month period from late January to early June in 2017. The hardware used is a Toshiba L500-1ZP laptop, which is a model from 2010 running a 64bit version of Windows 10. It has an Intel Core i3-330M processor running at 2.13 GHz and over its lifetime has been updated from 4 to 8 GB of RAM, and a solid-state disk with a total capacity of 232.89 GB of which roughly 50 GB were available. Considering the practical limitations and scale of the study area; to make the solar energy potential estimation manageable, the methodology involves at its core a sampling strategy. While data is available in the study area to perform a potential estimation similar to Google (2017), it was not possible with the available time, hardware, and scripting abilities of the author at the time to perform.

The chapter is structured as follows. First, the input data is discussed. Second, the sampling strategy. Afterwards the software used to model solar radiation and the validation process are described. Lastly, the scripting used to automate parts of the process is described.

3.1 Input data

One of the three main data inputs is a 2D building footprint dataset retrieved from the Danish geodata portal *Kortforsyningen*[21]. It is part of Kort10 which is a nationwide topographic object-oriented dataset in vector format at a scale of 1:10,000. It is maintained by the Data Supply and Streamlining Agency (*Styrelsen for Dataforsyning og Effektivisering*) which is part of the Danish Energy, Supply and Climate Ministry (*Energi-, Forsynings-, og Klimaministeriet*). The data is described as suitable for analytical purposes. Besides 2D building footprints Kort10 contains variety of data such as: administrative boundaries, infrastructure, nature (e.g. locations of water bodies and forests) and height curves.

The vector data is outlined with the use of photogrammetry applied to orthophotos which have a pixel resolution of 40 by 40 centimeters. Its metadata[22] mention that the accuracy is around 1m. The entire Kort10 dataset is updated every 5 years using the latest orthophoto, though traffic related data is updated three times a year. According to its metadata the version used in this thesis stems from May 2016. Therefore, any buildings that have been built, destroyed or altered after that time are not accounted for. The data is provided in ESRI shapefile format with an ERTS1989 UTM ZONE 32N spatial reference system and the

[21] Kortforsyningen: https://www.kortforsyningen.dk.
[22] Kort10 metadata: http://www.geodata-info.dk/Portal/ShowMetadata.aspx?id=49d0eee8-16c7-4e90-b91e-89fac91a892f

dataset with building footprints alone is 2.90GB in size when uncompressed. In total, it contains 4,567,813 building objects that represent:

> "… a **permanent** building, including garage, carport, tank/silo, outhouse, glass veranda, shed, any fixed covered areas, platform roof, houseboat and similar. But not mobile homes, campers, tents and similar."[23] (FOT, 2014, p. 16) (bold in original)

The dataset contains attribute values with a basic classification of each object being either a: building, houseboat, greenhouse, tank/silo, other, not reported or unknown. An attribute on the type of tank, which can be: closed, open, unknown, not reported and other. Moreover, an attribute indicating a status which can be: project approved, reported closed, reported up, under construction, under closure or in use. An additional attribute was added containing the area of each building footprint calculated in m^2 with a precision of four. A combination of these attributes is used to filter the dataset [ADD QUERY USED TO FILTER THE DATA]. In total 66,542 objects have been removed from the dataset with a total footprint area of 13,764,279m^2 or 13.7km^2. The total footprint area of all buildings and related objects in Denmark that remained part of the study is 696,615,648m^2 or 696.6km^2. The largest building (and related object) found in the dataset is 139,811.86m^2 and the smallest is 2m^2. The average size is 154.76m^2. Map 1 below provides an example of several footprints superimposed over satellite imagery.

Map 1 Building footprints superimposed over satellite imagery (source: Kortforsyningen)

[23] Translated from Danish: "BYGNING repræsenterer en permanent bygning, herunder garage, carport, tank/silo, udestue, glasudbygning, skut, halvtag, overdækket areal, fast markise på bygning, perrontag, husbåd og lignende. Men ikke mobilhome, campingvogn, telt eller lignende." (FOT 5.1 Specification, 2014, p. 16)

The second main data input are boundaries indicating types of urban development and land-use. They are also provided in the Kort10 dataset and are: city center, low-rise buildings, high-rise buildings, commercial/industry, mining sites, technical sites, horticulture, recreation, airports, sport facilities and cemeteries[24]. The specification mentions the existence of boundaries indicating public spaces, but these were not found in the Kort10 dataset at the time of writing (FOT, 2014, p. 211). Table 3 presents a description of class (FOT, 2014). This data is used to stratify the building footprints of which the justification and process is described in the next section. Map 2 shows an example of the different land use types overlaid on satellite imagery of a part of the city of Odense, located on the island of Fyn. These development and land-use boundaries are more detailed than CORINE and offer a higher differentiation within urbanized areas than what CORINE can offer. Therefore, these zones are therefore picked instead.

Map 2 Visualization of the urban-development and land-use classes (source: Kortforsyningen)

[24] Translated from Danish and presented in the same order as mentioned in the text: bykerne, lav bebyggelse, høj bebyggelse, erhverv/industri, råstofområde, gartneri, rekreativt område, lufthavn, sport anlæg, kirkegård, plads.

Table 3 Urban development and land-use classes

Urban development and land-use classes	*Description*
City center	Delineation of an area with coherent development in the center part of a city.
Low-rise urban development	Delineation of developed area with buildings up to two floors. Can include: freestanding housing, row housing, single family homes, apartment blocks, service oriented businesses, schools and institutions.
High-rise urban development	Developed area with buildings of more than two floors. Can include: apartment blocks, city blocks, service related businesses, schools and institutions.
Commercial/industry	Developed area with business related facilities in the form of: industry, crafts, shopping centers and industrial ports. It also covers all buildings, technical facilities, storage areas, parking lots and similar that belong to a business.
Technical sites	An area with technical functions or facilities. Can include: military areas, treatment plants, windmill parks, train stations, parking lots.
Airports	Delineates an airport and all related buildings and facilities in its vicinity.
Sport facilities	Delineates areas with facilities relating to sporting activities
Recreation	Areas used for recreational purposes with public access. Can include: parks, camping spaces, zoo's, botanical gardens and related.
Cemetery	Cemeteries and related facilities.
Horticulture	Facilities, buildings and areas used in the production of flowers, vegetables and fruits.
Mining sites	Sites of mineral extraction and related facilities and buildings.

The datasets with urban development and land-use boundaries is also preprocessed. Many polygons part of the same class seemed to overlap or overlap those of other classes. Filtering of these datasets was slightly less straightforward. The datasets contain an attribute indicating the time of their creation, but that did not make clear which one was still valid today and which was not. Another attribute that contained a special ID was used instead[25]. Those polygons within each dataset that had a value in this attribute were included, while those that didn't were excluded. This seem to have dealt with a lot of the within class overlap and part of the between class overlap, but not completely. This overlap is considered during the stratification of the buildings, discussed in the following section of 'Stratification, sampling and extrapolation'.

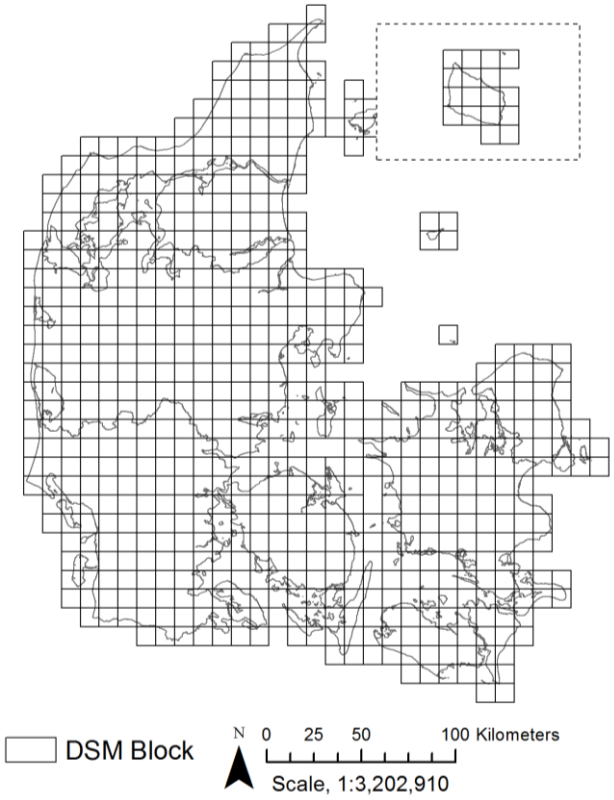
The third of the main data inputs is a DSM with a resolution of 40 by 40 centimeters, also retrieved from *Kortforsyningen*[26]. This is the most accurate DSM available for free and with its high resolution, it will be able to model any object that is equal or larger to 40 centimeters by 40 centimeters, meaning that it will help to distinguish the effects of existing roof uses and the impact they have on global radiation, for example. This particular dataset, is a resampled LiDAR based surface model of which the raw point data have been collected at an average amount of 4 - 5 points per m^2, by aircraft in the period of 2014 and 2015. The accuracy of this source data is 0.15 meters horizontally and 0.05 meters vertically. Currently, it can only be

[25] This attribute is called 'FOTFEAT_ID'.

[26] DSM metadata: http://www.geodata-info.dk/Portal/ShowMetadata.aspx?id=d4f542b3-7acb-4307-8589-f5c1f939ce00

retrieved in compressed zipped files representing 10km^2 blocks that each contain up to a 100 individual 1km^2 surfaces. The individual surfaces have an uncompressed size of 23.84MB and consist of 2500 columns and 2500 rows. They come in a GeoTIFF format with the same spatial reference system as the aforementioned vector dataset. Due to the time difference at which the DSM and vector data have been constructed, it is likely that there is a discrepancy between the two datasets that will provide some erroneous results across the study area. The author has not attempted to quantify this error, but it is expected to be marginal considering the scale of the study.

An additional vector dataset used in this study is a fishnet that specifies the location of the downloadable blocks, retrieved from *Kortforsyningen's* FTP server[27]. Map 3 shows this fishnet overlaid on the outline of Denmark. This fishnet is used to automate the process of retrieving the DSM blocks. An additional 1km^2 fishnet has been derived from the 10km^2 fishnet to automate the process of building a mosaic out of the individual 1km^2 surfaces that cover a sampling area. The automation process and sampling strategy is described in following sections.



Map 3 Grid locating downloadable DSM files

Several inputs are derived from the DSM that includes surfaces with slope, aspect (orientation) and horizon height values. These serve as input data for the solar radiation model, which itself outputs a new surface with solar irradiation values in Wh/m2 for an average day in a year at the same resolution of the

[27] *Kortforsyningen*'s FTP server: ftp://ftp.kortforsyningen.dk/

input DSM. An example of the latter is shown in figure 4 in section 'Solar radiation modelling'. Pixel-based statistics are calculated from the latter surface on a per building basis using the vector dataset of building footprints.

3.2 Stratification, sampling and extrapolation

To reiterate, the methodology here assumes that: Buildings of comparable size and in a like context will on average have similar levels of solar irradiation in Wh/m2/day on similar proportions of their rooftop. The dataset with building footprints does not provide information relating to roof structure, such as: size, height, slope, aspect and existing roof uses. The known variables, besides a count, are the shape of the footprint, size in m^2 and location. From the three available variables, the building size in m^2 and location are used. The former is positively related to the size of the roof. Therefore, it is assumed that footprint area is synonymous with rooftop area. The dataset with boundaries indicating urban development and land-use are used to stratify the buildings into different classes based on the location variable.

These development and land-use classes operationalize the aspect of context mentioned in the assumption, because these boundaries aim to characterize developed areas that are similar to one another in a consistent manner across the country of Denmark. An additional dependent variable used to operationalize context and help stratify the building dataset is building density. However, instead of density as a function of either building or population count, it is based on building footprint area in m^2. While the building count is the most independent variable available, since if there is no building there is no building footprint. The actual amount of rooftop area better reflects how densely built an area is compared to its count or population. For example, density will differ if there are 100 buildings with each a footprint of $125m^2$ in a defined area, or if there are 100 of them that have a smaller footprint. Stratification along this dependent variable is expected to be a suitable indicator of different urban structures, as it is derived from a variable directly related to the object under study: the building rooftop and its area.

The last aspect of the assumption: 'buildings on average have similar amounts of solar irradiation on similar proportions of their rooftop', is operationalized with the use of solar radiation modelling. Since the footprint dataset does not contain information on the roof structure, solar irradiation acts as a proxy of the heterogeneity of both the urban structure and roof structure that impact the solar energy potential of the object under study. It is used as such, because the literature on solar radiation modelling has shown that the amount of solar irradiance and irradiation under clear-sky conditions at the surface of the Earth (in this case rooftops), is highly depend on factors including: height, aspect, slope and neighboring objects that can cast shadows. In the process of modelling solar radiation on a sample area, the spatial variation of solar radiation on the individual rooftops will be able to indicate how heterogeneous the roof structures within that area. If radiation is uniform than roof structures are homogeneous and vice versa. A high variance will mean that an average will have trouble reflecting the heterogeneity among the roof structures.

The latter is tested in the end by comparing predicted global solar irradiation values in Wh/m2/day based on the averages found, against modelled global irradiation values in Wh/m2/day. This will indicate if the modelled solar energy potentials from a sample can reliably be extrapolated from a to the rest of areas part of a stratum.

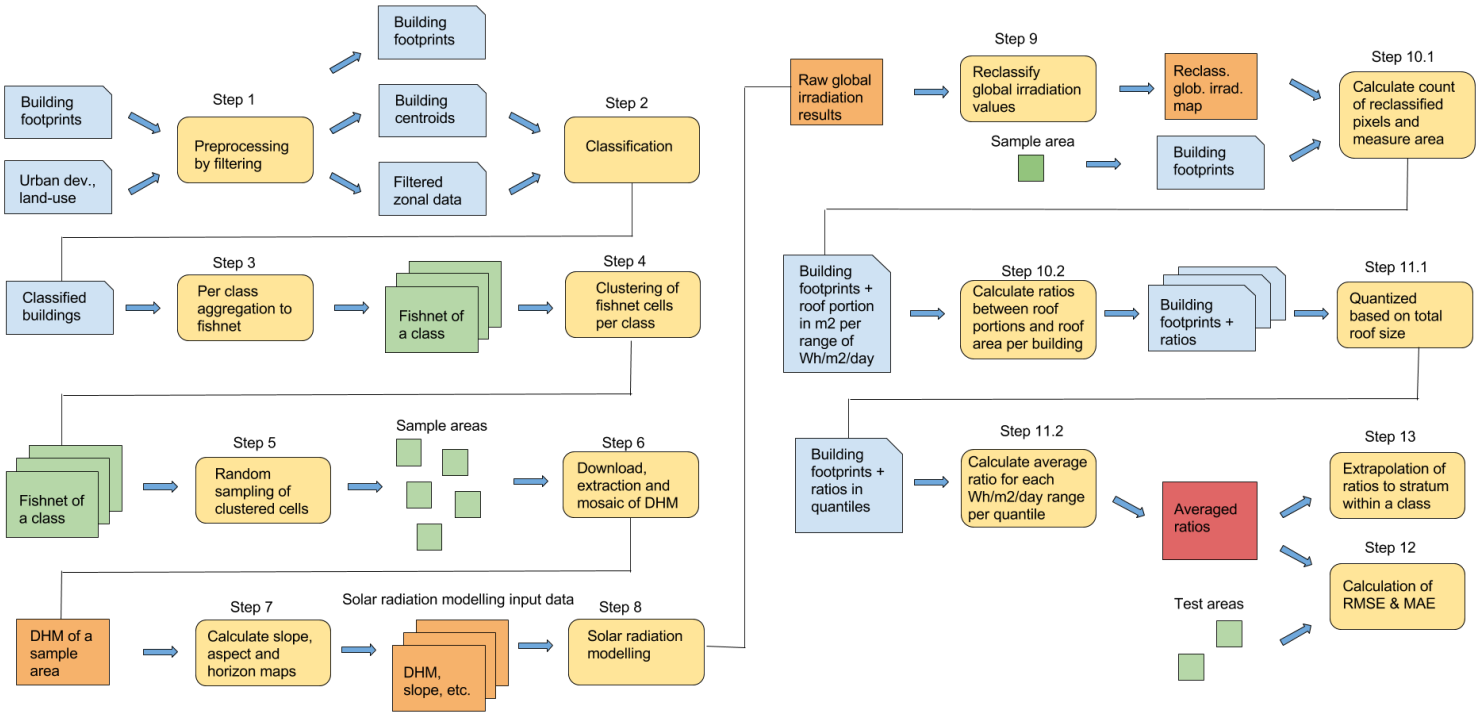


Diagram 1 Flow chart of applied methodology

Diagram 1 shows each step of the process part of the methodology. The process starts by preprocessing the building dataset and the urban development and land-use datasets and the conversion of the building 2D polygons into centroids. The point dataset is then stratified into the urban development and land-use classes based on a geometric intersection between the points and the polygon representing a class. Depending on the intersection a point is classified as part of one of the following: city center, low-rise buildings, high-rise buildings, commercial/industry, mining sites, technical sites, horticulture, recreation, airports, sport facilities and cemeteries. If a point falls within the boundaries of more than one class it will be classified as 'multiple' and an additional attribute will indicate what these classes are. Those points that do not fall within one or more classes are classified based on distance of which the procedure is as follows: The closest five boundaries are selected to point. The latter is then classified based on the shortest Euclidian distance between the actual footprint of the building that the point represents and a class boundary. An additional attribute is added to these points indicating how far the building was removed from the class it is classified as.

After the points are classified, a fishnet with a grid size of 2,500m by 2,500m is placed on top of them with an extent equal to the boundaries of the country of Denmark. The variable of rooftop size in m^2 is aggregated to this fishnet, along with a count of the amount of points. Both are added as separate attributes to each cell. This reflects a quadrat sampling method, which describes the process of sampling and aggregating point-based data inside consistently shaped regions (de Smith et al, 2015). Instead of using administrative boundaries to represent quadrats that have a varied sizes and shape, a uniform shape of a square is used to reduce the directional variance (Ibid, 2015).

The points are aggregated to a fishnet for several reasons. First, the solar radiation modelling is performed on an area, which is done because it allows the consideration of the surroundings of a rooftop and allow for precise modelling of clear-sky global solar radiation on the latter. It will likewise allow for the ability to extract results for multiple buildings at a time, on areas of which it is known that buildings are

present. Second, it would make the modelling manageable for practical reasons. Since if the buildings are sampled individually they would be scattered around the country of Denmark. Then multiple 10km^2 DSM blocks would have to be downloaded and the data handling would run into storage limitations. Third, it makes it possible to stratify the classified points, part of the same class, based on building density as a function of roof area in m^2 per 2.5km^2.

The roof area density is calculated by dividing the total roof area by the area of the cell. Afterwards, the cells of the fishnet are stratified along this density based on the univariate variance-minimization clustering method Natural Breaks/Jenks (Jenks, 1967). This algorithm aims to group data based on reducing the within group variance and maximizing the between group variance. It does not create equal sized groups and the groups are separated where there are substantial changes in values (de Smith et al, 2015). This clustering method is picked, because it will group areas together that have a similar roof area density and irrespective of their location in Denmark. The spatial location of the cells is not considered since it is believed that density reflects local conditions (Lehmann and Peter, 2003, Izquierdo et al, 2008). Therefore, when sampled it would mean that they are likely representative of each other.

There is an argument to base clustering also on location, since objects closer to one another are likely more similar to one another, than objects that are further away (Tobler's Law). On one hand, from the perspective of solar irradiation modelling, to take the location into account is good practice. If the study area consists of large latitudinal differences, an area in the south will have different insolation compared to an area in the north. However, Denmark is not very extensive. Its southernmost point is roughly at a latitude of 54.56 on the island of Falster, while its northern most point is roughly at a latitude of 57.75 at the tip of Jutland. Irradiation maps of DMI that indicate the average yearly sum of (real-sky) global radiation, do show a variation in global irradiation across the country of between 3,650-3,950 MJ/m2 (Wang et al, 2012, p. 9). However, the mentioned southernmost and northernmost latitudes both receive around 3,900Mj/m2. Therefore, if clear-sky global radiation is then considered (which is modelled here), then it is expected that global irradiation will not vary much between these latitudes.

On the other hand, out of practical reasons it also more manageable to cluster based on a univariate measure. Since, if both the location and roof density are considered, it would mean an exponential growth in the number of stratums and thus sampled areas that must be modelled. Since, per class, the fishnet cells would be grouped based on their proximity and afterwards differentiated by their roof area densities.

Currently, the amount of clusters a fishnet of a single class is stratified into based on roof area density, is grounded on an optimization method which is the Goodness of Variance Fit statistic (GVF) (Jenks, 1967). The GVF methods aims to minimize the squared deviations of the class means until the GVF is statistic is maximized at a value of 1 (Ibid, 1967). The threshold of this statistic is set at a value of 0.9 and the number of clusters created increases until this threshold is met -although a hard cap is set at 5. Afterwards, per cluster two groups of five random cells are selected. Their representativeness of the cluster is tested based on a parametric dependent T-Test when the distribution of roof density of the cluster is normal, or a non-parametric Wilcoxon signed rank test is if this distribution is not normal. Normality of the distribution of the cluster is tested using a normality test. Both them will test if the two dependent groups of five random fishnet cells have been taken from the distribution of the cluster by comparing the mean values. If the means of the samples significantly differ from the mean of the cluster, then they are not representative and two new random samples are picked and tested again. The significance level is set at 0.1. If they pass the test the first of two groups are exported and used as sampling areas for solar radiation modelling. If a cluster is smaller or equal to five, the entire cluster of fishnet cells is taken as the sampling areas.

For each of the five sample areas, a DSM is constructed from the 1km^2 DSM surfaces that overlap the sampling area with the addition of a buffer of 600 meters in every direction. The DSM surface is larger than the sampling area to make sure that buildings located along the edges of a sample area are properly evaluated against their context. Solar radiation itself is only modelled on the sampling area plus 100 meters in each direction to make sure that the entirety of the roof along the edge of a sample area is evaluated.

After solar irradiation is modelled on a sample area in Wh/m2/day, the values are reclassified into seven ranges: 0 – 1000, 1000 – 2000, "…", 5000 – 6000 and 6000 and above. With the use of raster statistics, the number of pixels per range of Wh/m2/day are counted per roof outline. The count per range is multiplied by 0.16 to measure its size in m^2. The area in m^2 of each range, as they are present, are divided by the total size of the roof, measured by multiplying the total count of pixels within a roof outline by 0.16. The roof size in m2 is based on the count of pixels, since the roof outlines themselves are line-based while the pixels are squares that only approximate to this line. By dividing the two the ratio of the rooftop area that received a range of solar irradiation is given. The area in m^2 of the entire roof based on pixels, the area in m^2 of each portion and the ratios for each portion are added as attributes to the rooftop outlines

Lastly, all individually evaluated buildings across all five sample areas, are divided into quantiles based on their roof area in m^2 (derived from the pixel count). Then for each quantile, the ratios are averaged. This results in seven averaged ratios per quantile. These averaged ratios are the coefficients used to calculate the average solar energy potential of all buildings within the same quantile range, of those fishnet cells that are part of the same cluster.

3.3 Validation

To test the performance of the coefficients a test sample is taken from the same cluster, for which the process up to the calculation of the ratios per individual rooftop are performed as well. Then, the averaged ratios per quantile of the sampled areas are applied to those of the test sample that fall in the same quantile range, to predict the area in m^2 on a rooftop that receives a range of Wh/m2/day. These predictions are compared to the test samples own modelled results, by calculating the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) between the distributions of each predicted and measured area per range of Wh/m2/day. The lower the MAE and RMSE values the better the coefficients perform. This will indicate if the applied stratification is able to consider the heterogeneity of roof structures appropriately and gives insight into how much the prediction is of for that stratum providing more authority to the results.

3.4 Modelling of clear-sky global solar radiation

The solar radiation model used is an adapted version of rsun implemented in the GRASS GIS platform, called r.sun.daily[28]. It is developed by the NCSU GeoForAll Lab located at the North Caroline State University, Center for Geospatial Analytics in Raleigh, NC, USA[29], who also developed the GRASS GIS platform (Neteler and

[28] r.sun.daily manual: https://grass.osgeo.org/grass72/manuals/addons/r.sun.daily.html
[29] NCSU GeoForAll Lab https://geospatial.ncsu.edu/osgeorel/

Mitasova, 2008). The main difference is an additional parameter that allows the user to parallelize the modelling process across multiple processors. The result is that multiple maps can be calculated at the same time depending on the number of processors the user has available, which reduces the computational time by multiple factors. The rsun model is picked over the Solar Analyst model as it allows for a more flexible use of input data and optimization, due to the source code being open. The latter also allows it to be incorporated into automated processes such as is presented in this thesis and which is examined in section 'Scripting and reproducibility'. In addition, it would allow for reproducibility of the results by anyone since the Solar Analyst model is only available in the commercial GIS platform of ArcGIS.

The solar radiation that is modelled is a daily sum of clear-sky global irradiation in Wh/m2/day for the center day of each month within a year. The twelve resulting rasters are summed and divided by twelve, to give an averaged irradiation value representative of an average day in a year. Irradiation is not modelled for every day of the year and then divided by 365, due to storage and time constraints. By using the center day of each month, it would roughly approximate to the same result in clear-sky global irradiation in Wh/m2/day. Real-sky solar irradiation is not modelled. This will mean that the modelled radiation values are overestimated compared to reality, since they are not modified by the impact of clouds. However, clear-sky global solar radiation will be better suited to distinguish between rooftop space that has a higher solar energy potential on average – due to for example, existing roof uses or nearby objects casting shadows or due to its inclination and orientation. If real-sky irradiation was modelled this distinction would be more difficult to make since at times the diffuse component would dominant that reduces the impact that for example slope and orientation have. Although an approximation of real-sky irradiation would be necessary to predict the productivity of a roof in terms of an applied solar energy conversion technology. If clear-sky irradiation values are used this will over estimate productivity values.

The input data used are:

- A DSM of 40x40 centimeters with an extent of 3.7 by 3.7 kilometers
- A slope and aspect raster of the same size as the DSM
- A buffered 2D polygon representing the sample area of 2.7 by 2.7 kilometers
- 24 pre-calculated horizon maps of 3.7 by 3.7 kilometers.

The horizon rasters are calculated with the use of a separate and modified r.horizon script[30], which like r.sun.daily allows for parallelization of the calculations across multiple processors. 24 of them are calculated equal to the number of hours in a day for which solar irradiation is modelled before it is summed to a daily sum of clear-sky solar irradiation in Wh/m2/day. The 2D polygon sets the extent of the area on which solar irradiation is modelled, which is slightly larger than the sample area itself to make sure that roof of buildings on the edges are evaluated properly. Other parameters that are set within the r.sun.daily model include a linke turbidity factor, that models the transmissivity of the atmosphere (set at the default value of 3.0) and a ground albedo factor to model the reflectance of the surface (also kept at a default of 0.2). The 2D polygon sets the extent of the area on which solar irradiation is to be modelled.

[30] Link to source of modified script: http://osgeo-org.1560.x6.nabble.com/Simultaneous-r-horizon-processes-td4676325.html#a4892854. This script was further adapted to suit the needs of the author.

3.5 Scripting and reproducibility

The process described in the section on stratification, sampling and extrapolation, and validation is automated with the use of scripting in the 32bit version of the Python 2.7 programming language. Several open source python libraries were utilized in the process which are described in Table 4. The dependencies of these libraries are not mentioned, as well as those that are already built-in. Solar radiation modelling was performed within the GRASS GIS platform which is a 64bit environment to be able to utilize all available memory during the modelling process. Most of the packages have been retrieved from a repository of unofficial Windows binaries for python extension packages, managed by the Laboratory for Flourescence Dynamics at the University of California, Irvine[31]. They are pre-compiled and make matching versions of python with its respective extension packages more accessible.

Table 4 Utilized Python extensions and libraries

Library	*Version*	*Description*	*Source*
GDAL	2.1.3	Data handling and basic analysis of both raster and vector based data	http://www.gdal.org/
Rtree	0.8.3	Library to build spatial indexes on data to improve query speeds	https://pypi.python.org/pypi/Rtree/
Jenks	1.0	Library used to perform Jenks variation-minimization clustering	https://github.com/perrygeo/jenks
Numpy	1.11.3	Package used for scientific computing with Python	http://www.numpy.org/
Rasterstats	0.12.0	Library to perform summary statistics on rasters and vector geometry	https://github.com/perrygeo/python-rasterstats
Scikit-learn	0.18.1	Data mining and analysis library	http://scikit-learn.org/stable/
Scipy	0.19	Python-based ecosystem of open-source software for mathematics, science, and engineering	https://www.scipy.org/
Grass.script	7.2.1	Python library included in GRASS GIS	https://grass.osgeo.org/

With the use of these Python extensions a multitude of scripts were made to handle the steps presented in Diagram 1. In table 5 each script is presented and which libraries they utilized including those that are built into Python 2.7. In the 'Step' column a number is provided that matches the processing step in Diagram 1. These scripts are not directly applicable on another system in their current state. For example, they would have to be edited to make sure that the data handling aspects -required file folders and naming schemes- work correctly. In addition, it would have to be made sure that a similar version of Python and extensions is installed. A great deal of streamlining and efficiency gains can still be made in the scripting of this process. However, this could not be achieved within the available time. The developed scripts are available for download[32].

[31] Repository: http://www.lfd.uci.edu/~gohlke/pythonlibs/

[32] Download link to scripts: https://drive.google.com/file/d/0B_kmPxXGEcYDLVpHR29ISVFLdE0/view?usp=sharing. Note: the scripts are very much in a development stage, which means that they lack commenting and proper structure.

Table 5 Python scripts

Script	*Step*	*Libraries and modules utilized*
Building preprocessing	1	GDAL
Class preprocessing 1	1	GDAL
Class preprocessing 2	1	GDAL
Building categorization	2	GDAL, Rtree, collections, operator
Sampling fishnet	3	GDAL, math
DSM fishnet preprocessing 1	3	GDAL, Rtree, math
DSM fishnet preprocessing 2	3	GDAL, Rtree, itertools
Centroid aggregation	3	GDAL, Numpy
Clustering preprocessing	4	GDAL, Numpy
Clustering process Jenks	4	GDAL, Numpy, Jenks
Random sampling	5	GDAL, Numpy, Scipy, random, itertools
Mosaic DSM	6	GDAL, Numpy, zipfile, subprocess, glob, ftplib, operator, itertools
Grass import 1	7	Grass.script, GDAL
Grass import 2	7	Grass.script, GDAL, zipfile, numpy, subprocess, glob, operator, itertools
Grass horizon	7	Grass.script, multiprocessing, datetime
Grass rsun	8	Grass.script, gdal, multiprocessing, datetime
Extraction preprocessing	9	Manual
Extraction factors 1	9	GDAL, Numpy, rasterstats, subprocess, csv
Extraction factors 2	10.1, 10.2	GDAL, Numpy, rasterstats, subprocess, csv
Calculate averaged ratios	11.1, 11.2	Manual
Validation quantiles	12	GDAL, Numpy
Validation errors	12	Numpy, Scikit-learn
Extrapolation	13	GDAL, Numpy

4. Results

4.1 Stratification, aggregation and clustering

Table 6 shows the stratification of the buildings up to the classification into urban-development and land-use types. It indicates the class name, the count of buildings, percentage of the total count, average roof area in square meters, total roof area in square kilometers and percentage of the latter's total. To reiterate the numbers are based on building footprint area which are assumed synonymous with roof area. The results indicate that most buildings are situated within low-rise urban development (92.86%). Followed by commercial/industry (2.93%), city center (1.50%) and high-rise urban development (1.46%).

Table 6 Building classification results

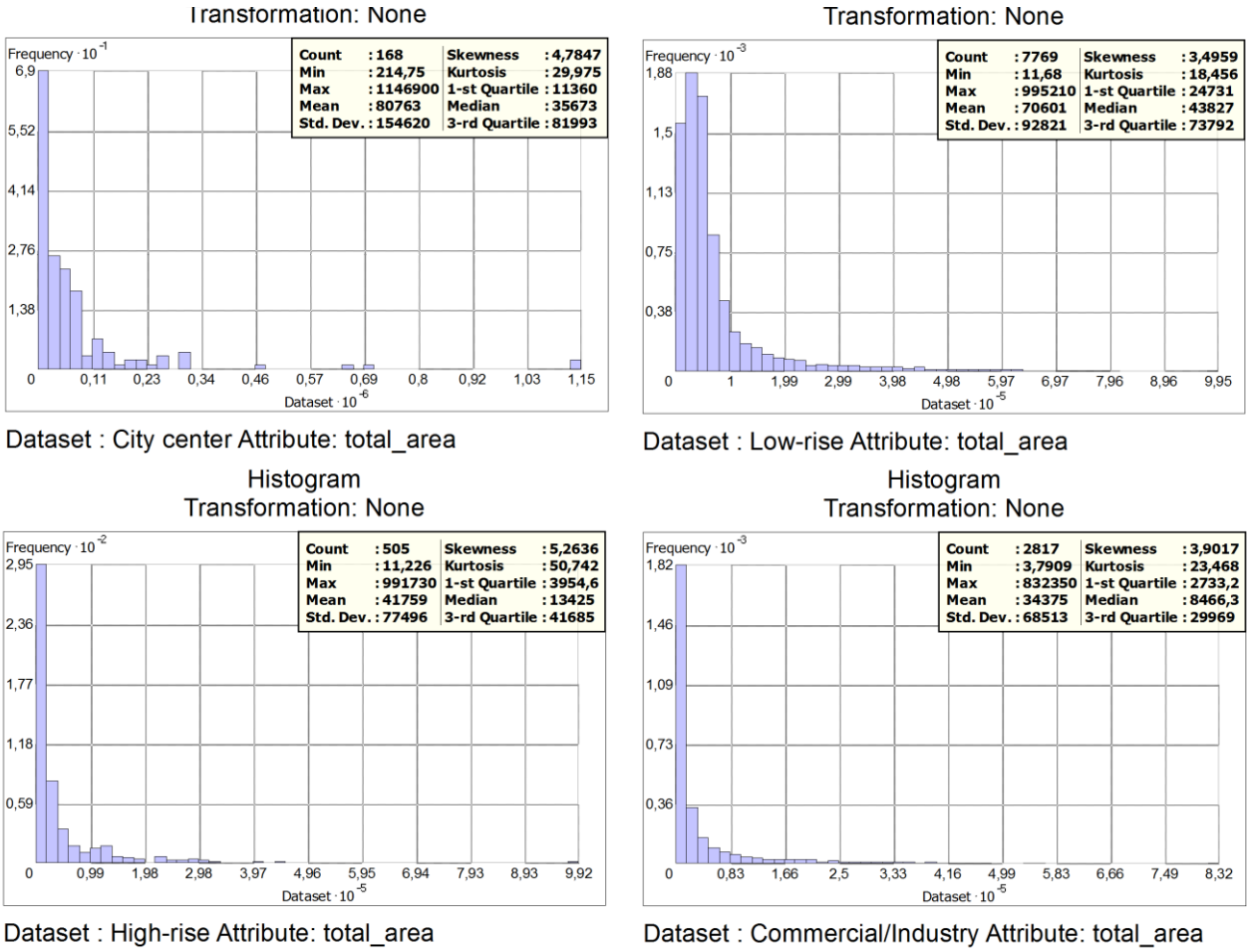
Class	*Count*	*Percentage of total count*	*Average roof area in m^2*	*Total roof area in km^2*	*Percentage of total roof area*
City center	67,899	1.5084	199.83	13.57	1.9479
Low-rise buildings	4,180,052	92.8638	131.23	548.50	78.7374
High-rise buildings	65,890	1.4638	320.05	21.09	3.0275
Commercial/industry	131,934	2.9310	733.96	96.83	13.8999
Mining sites	813	0.0181	161.53	0.13	0.0187
Technical sites	17,886	0.3974	280.89	5.02	0.7206
Horticulture	4,295	0.0954	1045.83	4.49	0.6445
Recreation	2,285	0.0508	56.62	0.13	0.0187
Airports	1,583	0.0352	629.36	1.00	0.1436
Sport facilities	14,540	0.3230	287.42	4.18	0.6000
Cemeteries	4,853	0.1078	149.76	0.73	0.1048
Multiple	9,241	0.2053	102.52	0.95	0.1364
Total	4,501.271	100	154.76	696,62	100

Some of the other classes have a significant count (technical sites and sport facilities) but do not contribute much to the total in terms of roof area in Denmark (less than one percent). Besides the low-rise development, the commercial/industry class sets itself apart. It contains roughly double the amount of buildings compared to city center and high-rise development, but has five to seven times the amount of roof area and 13.89% of the total roof area. Low-rise urban development is most common in Denmark and has 78.73% of the total roof area. On average, buildings in the horticulture class have the largest roof space followed by commercial/industry, airports and high-rise urban development.

Chart 1 contains four histograms with the distribution of total roof area in m^2 aggregated to a fishnet of 2.5km^2 cells of the four largest classes. The data is grouped in fifty bins, the units are m^2 and only those cells with a building count above zero are included. The count indicates that not all classes are present in the 8,179 fishnet cells covering the entire study area. City center is only present in 168 cells, whereas high-rise is present in 505 and commercial/industry in 2,817 cells. Low-rise is present almost everywhere with

7,769 cells having one or more building in them. The distribution of all four largest groups is heavily skewed with the tail to the right, thus in each class several cells contain considerably more roof area compared to the majority. Additional descriptive statistics are included with each histogram with information on for example: the minimum and maximum total roof area, and the average total roof area.

Chart 1 Histograms of the distribution of total roof area of the four largest classes of buildings in Denmark



Map 4 contains four maps showing the spatial distribution of total roof area aggregated to the fishnet cells of the same classes. The data is visualized using quintiles based on the data of each class separately. Again, only those cells of the fishnet that contain at least one building are visualized and we see that not all classes are present everywhere in the study area, except for low-rise development. This class is spread across the country in varying degrees with the largest concentrations found in the main cities and their suburbs. High-rise is mostly found in larger cities coinciding with the spatial distribution of the city center class in the top left. Commercial/industry is similarly to low-rise development quite spread out, but is most concentrated near the cities. The island of Bornholm, indicated in the top right of each map, shows similar variations per class as the rest of the country.

Map 4 Spatial distribution of total roof area across Denmark

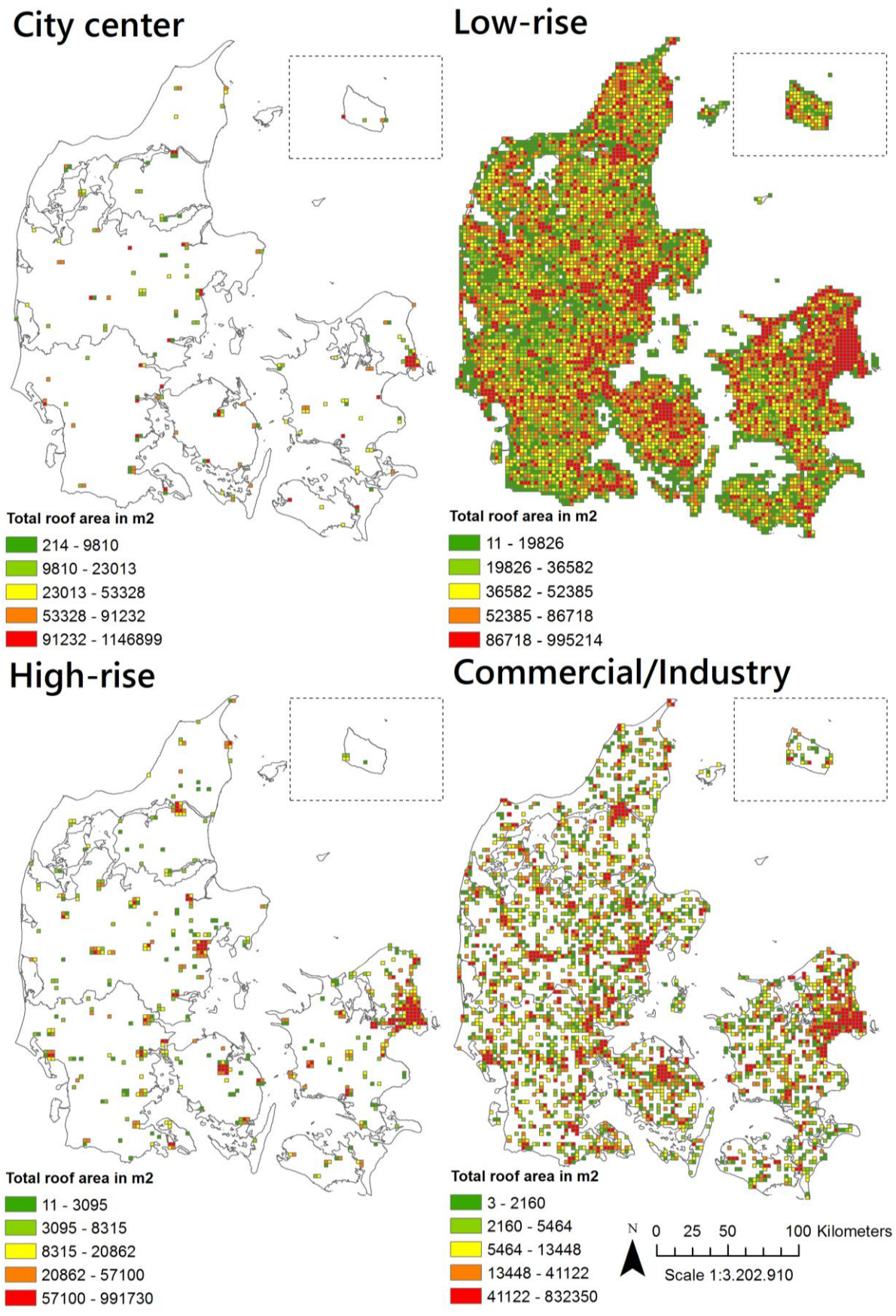
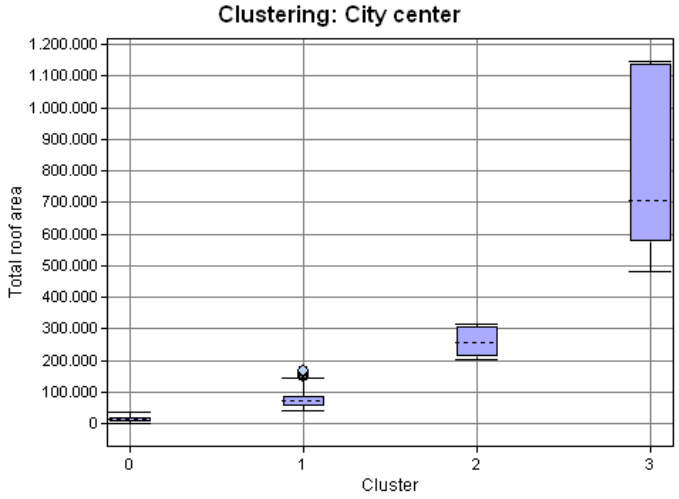


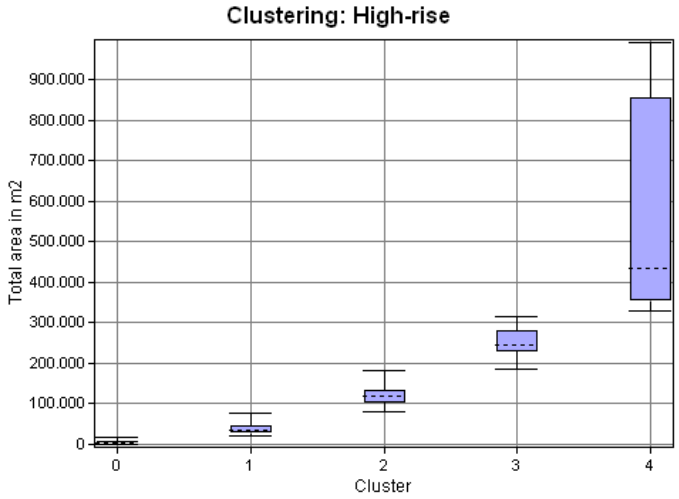
Chart 2 show four graphs with boxplots of the results of the univariate clustering process using the Natural Breaks Jenks algorithm. The data of the four largest classes are displayed and the x-axis specifies the cluster from low to high, which reflects low to high density in terms of total roof area m^2 on the y-axis. With the use of this algorithm, groups of different densities have clearly been delineated. Particularly, if the interquartile ranges between the clusters are compared, specified by the light blue boxes. However, the highest cluster in each class has a relatively higher variance compared to the others. City center and high-rise especially, of which the interquartile range ranges from roughly 600,000m^2 to 1,150,000m^2 for the former, and roughly 360,000m^2 to 850,000m^2 for the latter. Although, the count of cells contained in these clusters is comparatively much smaller, indicating that they are a few outliers with relatively high density that have been clustered together. Moreover, not all classes displayed here have the same number of clusters. Both city center and commercial/industry met the Goodness of Variance Fit (GVF) threshold with 0.906 and 0.917 respectively, with just four groups. Low-rise reached the threshold with five clusters (GVF: 0.931) but high-rise was exported as five clusters with a GVF of 0.899. It met the maximum of five clusters without reaching the thresholds (though only by a little). In sum, the algorithm performs well in stratifying the classified buildings further into several clusters with a reduced variance of total roof area in m^2.

Chart 2 Boxplots of clustering results per class for the largest classes

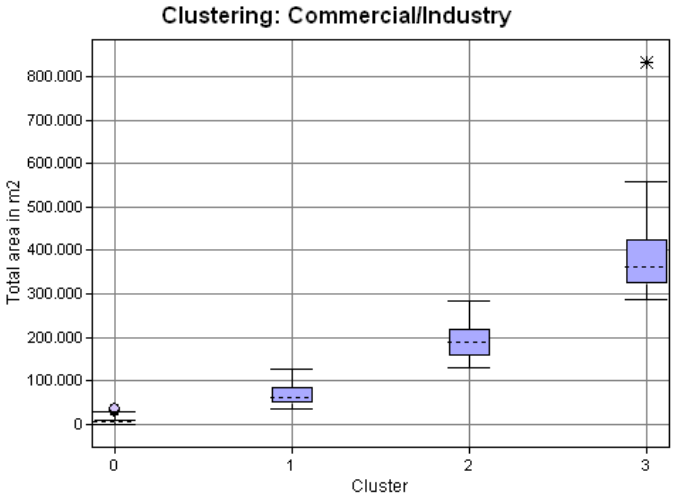


Count: Cluster 0: 86, Cluster 1: 65, Cluster 2: 12, Cluster 3: 5

Count: Cluster 0: 3844, Cluster 1: 2827, Cluster 2: 704, Cluster 3: 281, Cluster 4: 113



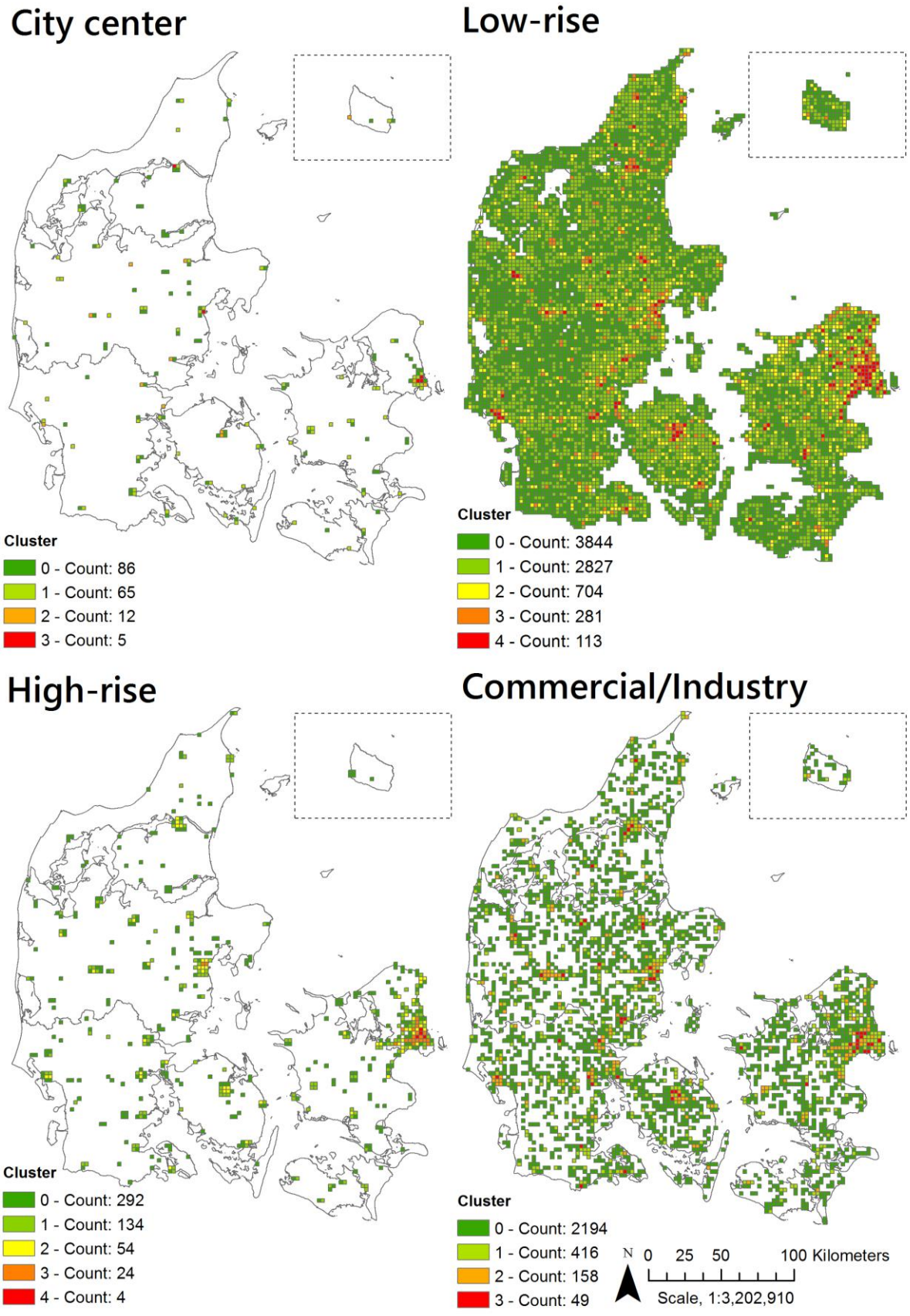
Count: Cluster 0: 292, Cluster 1: 134, Cluster 2: 54, Cluster 3: 21, Cluster 4: 4



Count: Cluster 0: 2194, Cluster 1: 416, Cluster 2: 158, Cluster 3, 49

Map 5 (on the following page) indicates the spatial distribution of the clusters across Denmark and as expected, the higher density clusters of each class are near or in the larger cities of the study area. City center can obviously only be found in the larger cities, with the highest densities found in the cities of Arhus (East Jutland), Aalborg (North Jutland) and Copenhagen (North-East Zeeland). The highest density cluster 3 and 4 of the high-rise class are mainly found in Copenhagen, while the same high-density clusters of the low-rise class are more spread out with concentrations in other cities as well. Commercial/industry follows a similar trend to the low-rise category with low-density cells being more numerous and spread across the country compared to the higher density clusters with less numerous cells.

Map 5 Spatial distribution of the clusters per class for the largest classes

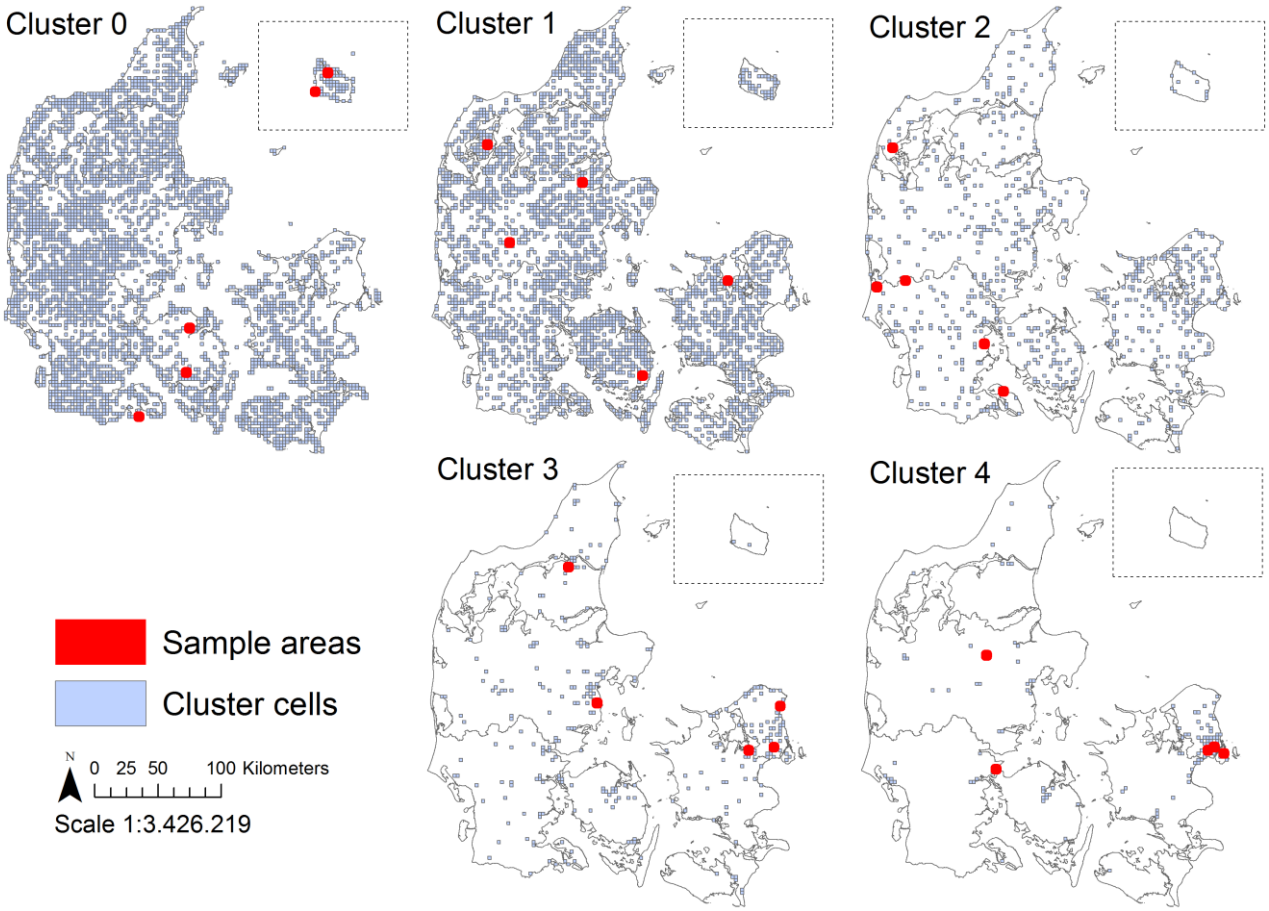


4.2 Sampling results

This section describes the sampling results. In the automated sampling process ten samples are taken from each cluster at random and tested for their representativeness against the cluster they are taken from. Their representativeness was tested with the use of either a dependent T-Test or Wilcoxon signed-rank test, with the significance level of p = 0.1. If the tested sample's p value is higher a new sample is picked and tested. If their p value is below 0.1 the first of the two groups of randomly picked samples are exported and used as sampling areas to model solar radiation on.

Map 6 shows the location of the samples taken from each cluster compared to the rest of the same cluster for the low-rise urban development class. It gives an impression of how the random sampling functions. As expected, due to the samples being randomly picked, they may not be evenly spread from the perspective of spatial sampling. The argument for not taking this into account is presented in the methodology chapter. Moreover, each cluster is represented by the same number of samples, regardless of how numerous the cluster is. These samples passed the dependent T-Test or Wilcoxon signed-rank test. Although, from a statistical point of view the number of samples may not provide results with enough confidence. However, out of practical reasons the number of sampling areas had to stay low, which is elaborated in the methodology chapter.

Map 6 Example of random samples taken per density cluster of the Low-rise class



Due to a lack of time only the samples that come from Cluster 0 of the low-rise urban development class presented in Map 6 have had their clear-sky solar radiation modelled. Therefore, these sampling areas are examined in more detail here, before moving on to the solar radiation modelling results. First, is a table on the sample areas themselves. Table 7 shows the ID of the cells, their density, count of buildings contained within the cell and total, average, minimum and maximum roof area in m2. They are the same cells as depicted in Map 6 and are comparatively of low density, though they represent 3,488 cells that are part of the same cluster. Interestingly, while there is a positive relationship between count and total roof area, here we can see that the total roof area does not linearly increase with count. Therefore, this is an indication that total roof area better reflects density compared to count, as mentioned in the methodology.

Table 7 Descriptive statistics of sampled areas

Sample ID	*Density*	*Count*	*Total roof area*	*Avg. roof area*	*Min. roof area*	*Max. roof area*
6051	0.0067	225	42,178	187	2	2,327
8152	0.0056	232	35,092	151	6	1,250
8279	0.0049	249	30,697	123	2	1,488
23801	0.0004	31	2,545	82	11	313
24359	0.0035	101	22,085	219	17	1,262

Chart 3 Distribution of roof area size in m2 per sample and their location

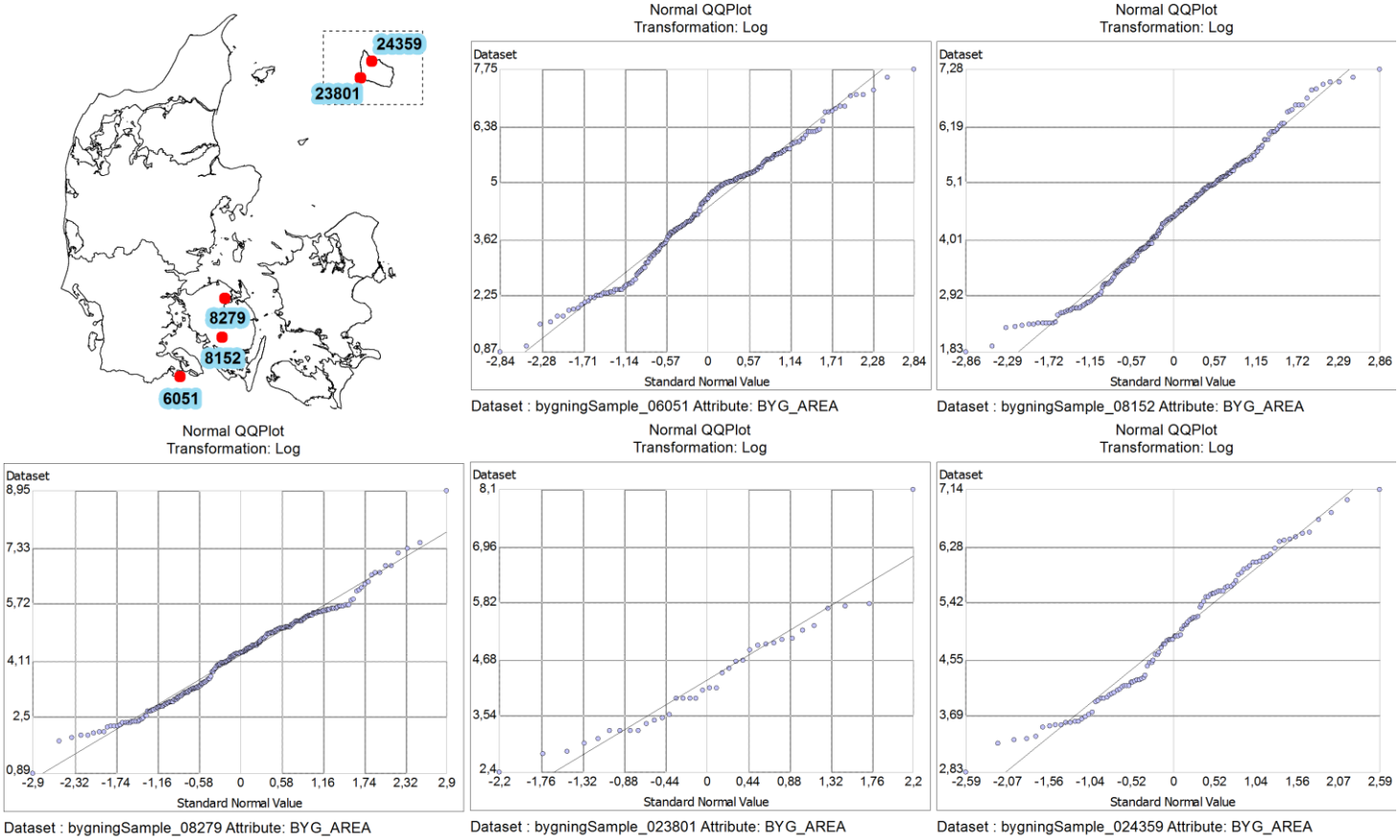


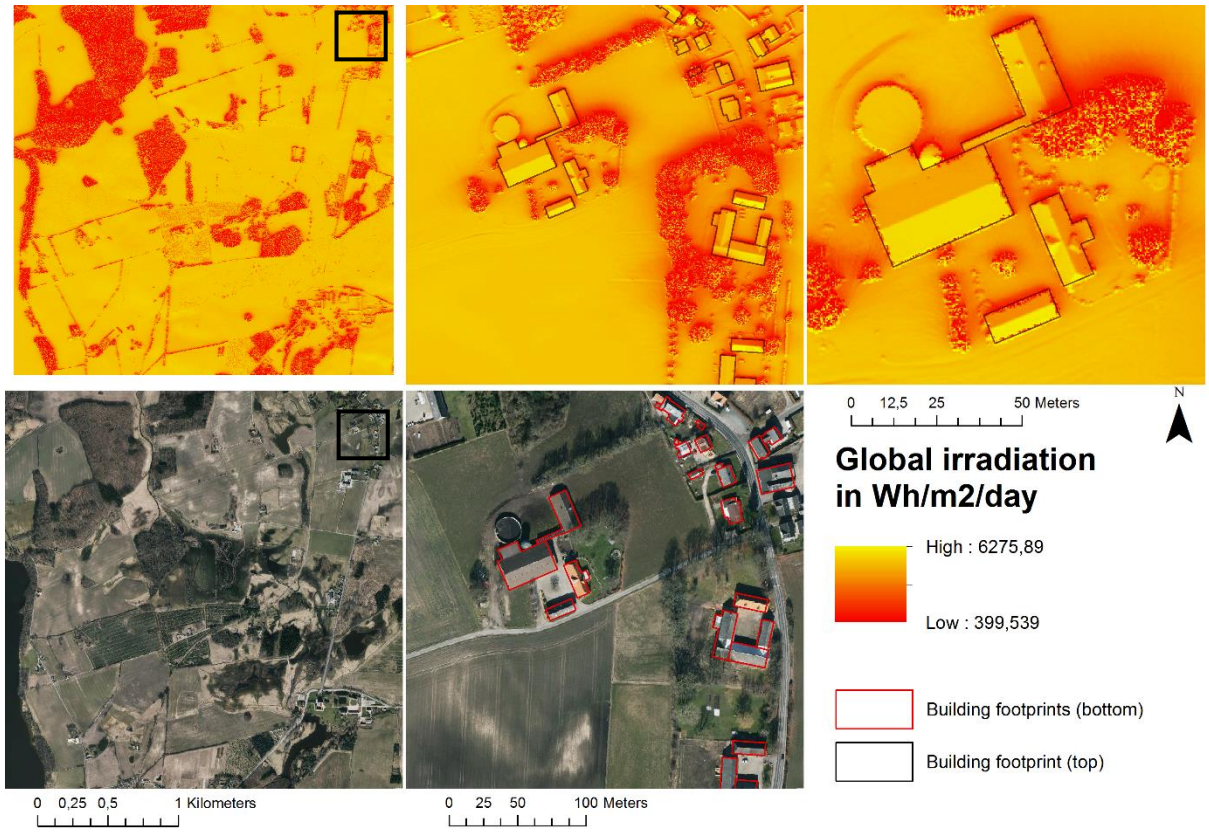
Chart 3 presents locations of each sample area and Normal QQ plots of the distribution of roof area in m^2 of the buildings present in each sample area. The distributions are presented in logarithmic scale to make the distributions more interpretable and we can see that they are have a lognormal distribution with a slight variation in steepness. The highest and lower values tend to deviate slightly from it.

4.3 Results of solar radiation modelling

In Map 7 at the bottom of this page, the modelled clear-sky global irradiation results on the sampling area with ID '8152' is depicted. On the top row is the raw output visualized on a continuous scale. Lower values have a darker red color, while higher values move towards a bright yellow. In the top left image, the modelled area is shown in its entirety. In the middle, it is zoomed in to the boxed area and the footprints of the buildings are included delineating roof area. The image in the top right zooms in even closer on a small group of buildings, found in the middle image. Below the top row are two orthophoto's taken in spring of 2016 of the same areas to give an indication of how they look in reality.

With the use of the 40 by 40 centimeters DEM it is possible to model a detailed spatial variation of the clear-sky global irradiation across rooftops and areas in between. The effects of rooftop orientation, slope and shadowing caused by existing roof uses, as well as neighboring objects are clearly visible. Thus, it is possible to distinguish rooftop area that consistently receives more radiation. In the top-right image, we can see these effects most clearly. The irradiation values indirectly distinguish different roof portions, their orientations and slopes, and existing roof uses based on the effects these factors have on clear-sky global irradiation. For example, the roof of the upper most building is shadowed partially by the adjacent tree-line.

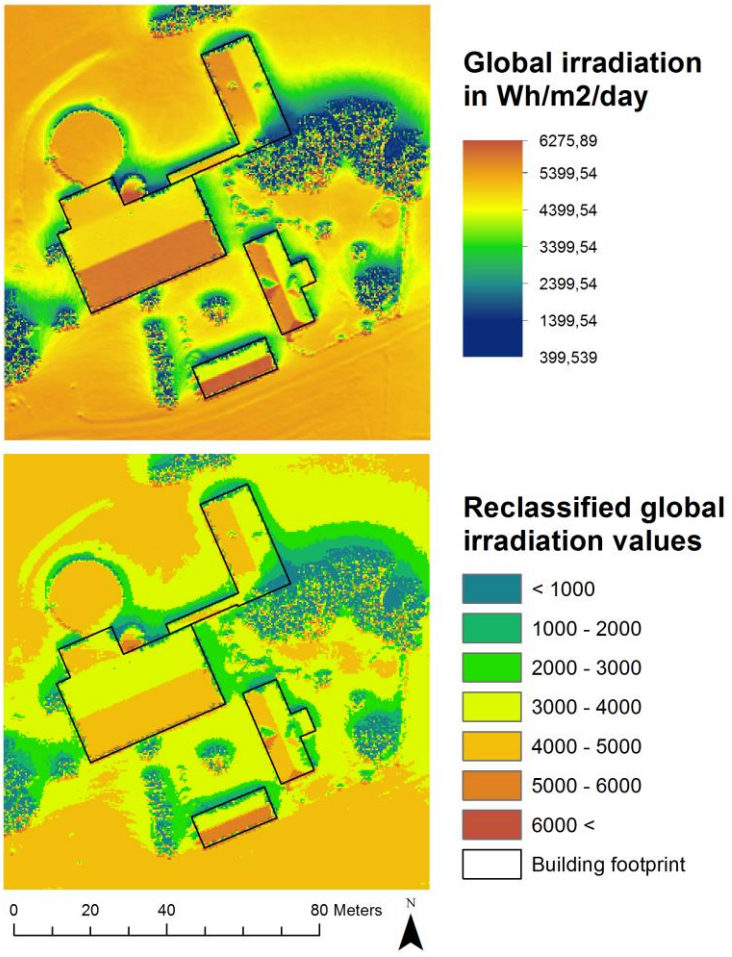
Map 7 Solar radiation modelling results for Sample area with ID 8154



The large rectangular building in the middle has a substantial portion of its roof facing south-east with an unobstructed view of the sky, as it has the highest values of irradiation depicted by the bright yellow.

The raw global irradiation values in Wh/m^2/day have been reclassified into seven ranges (< 1000, 1000 – 2000, "…", 5000 – 6000, 6000<). In the process of reclassifying the raw values, the spatial variation becomes slightly generalized, as shown in the bottom image of Map 8 below. But the most detail is preserved allowing for the distinction of different portions of the roof.

Map 8 Reclassification of modelled solar radiation values

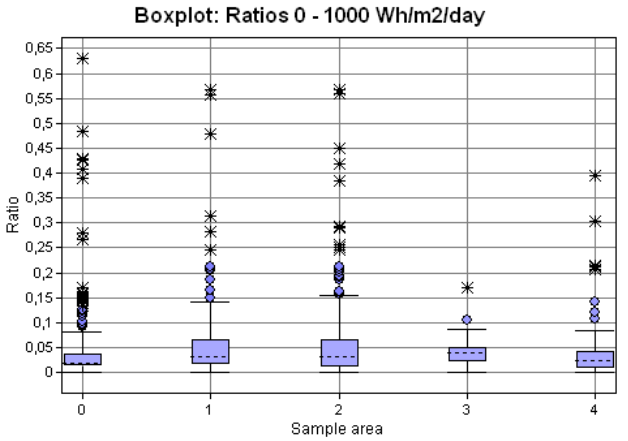


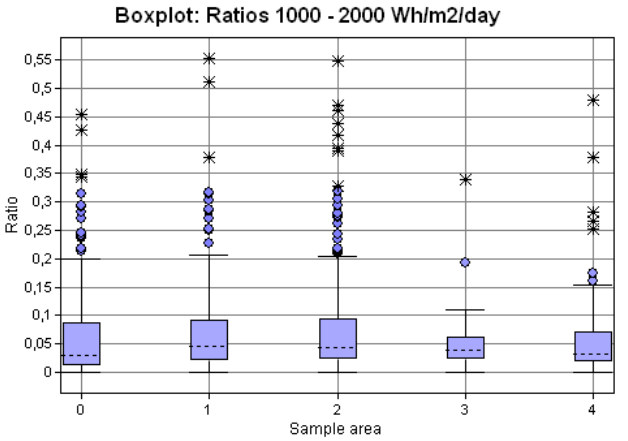
Within each of the building footprints, pixels with a value that fall in the specified ranges, are counted and multiplied by 0.16. This gives the portion's area in m^2 and depending on the number of ranges that are present on a rooftop, there are up to seven portions per roof. Per building, these portions are divided by the total roof area represented by pixels to get the portions ratio. In chart 4 and 5, the distributions of each ratio per sample area are presented in boxplots to show the spread.

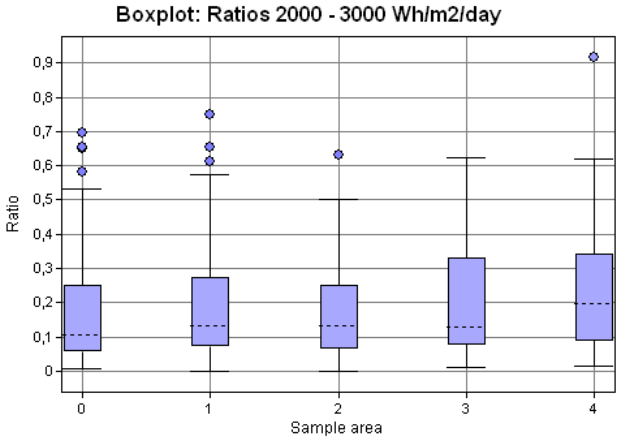
Irrespective of building size in m^2, it can be seen that for the interquartile ranges of each ratio across the sampling areas, that there is a general increase in the proportion of roof covered by a range, as the range increase up to 4000-5000 Wh/m2/day and 5000-6000 Wh/m2/day. The ratios decrease for the range of 6000 Wh/m2/day and above. However, at the same time, the variability also increases as can be

seen by the increasing length of the boxes and the whiskers. The exception is the range between 5000-6000, where the right most boxplot has a much larger variance compared to the other sampling areas. Moreover, in almost all cases there are roofs where a range is not present at all, shown by the bottom whiskers stretching to a value of 0.

Chart 4 Boxplots of ratios per sample area per range of Wh/m2/day (part 1)







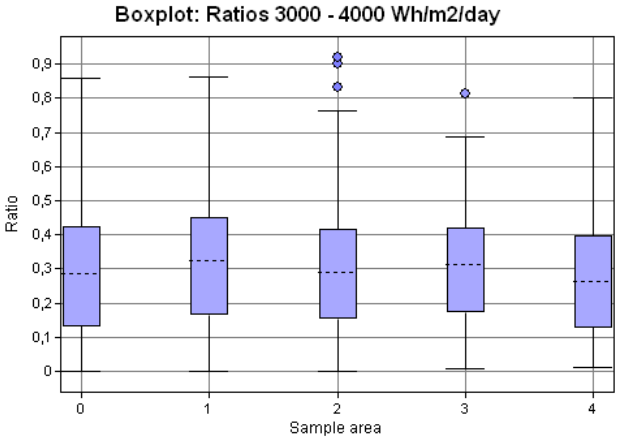
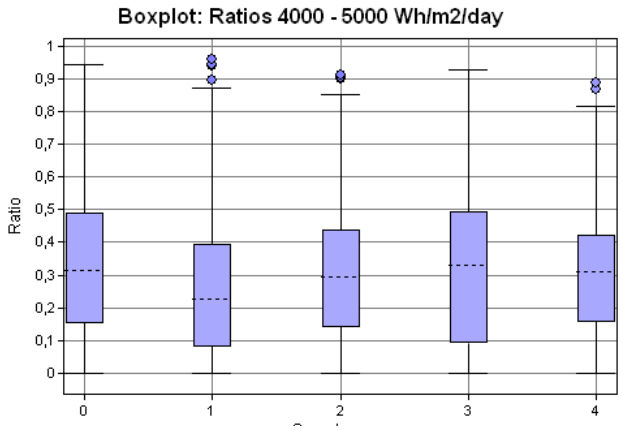
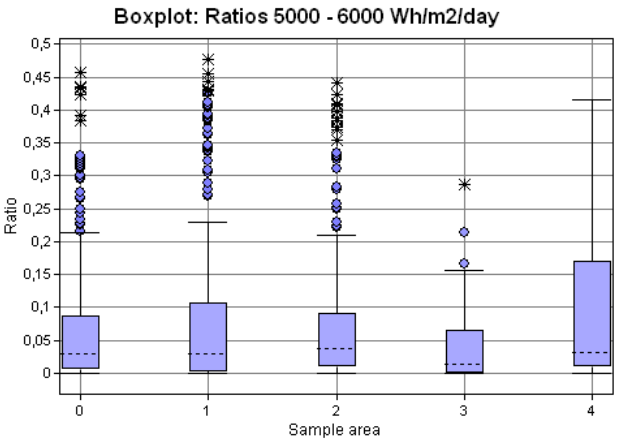
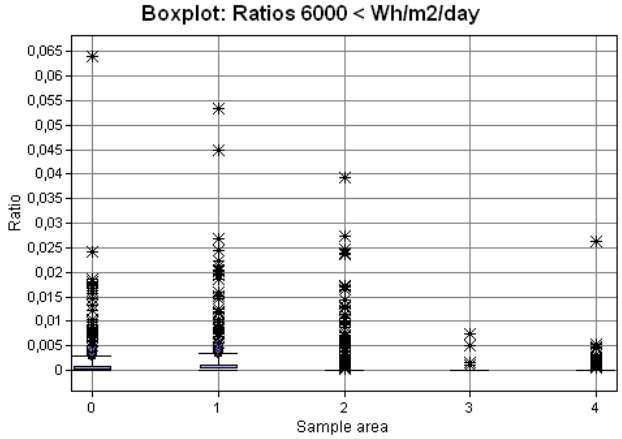


Chart 5 Boxplots of ratios per sample area per range of Wh/m2/day (part 2)





Boxplots: 0 - Sample 6051, 1 - Sample 8154, 2 - Sample 8279, 3 - Sample 23801, 4 - Sample 24359
Counts: 0 - 225, 1 - 238, 2 - 265, 3 - 36, 4 - 105



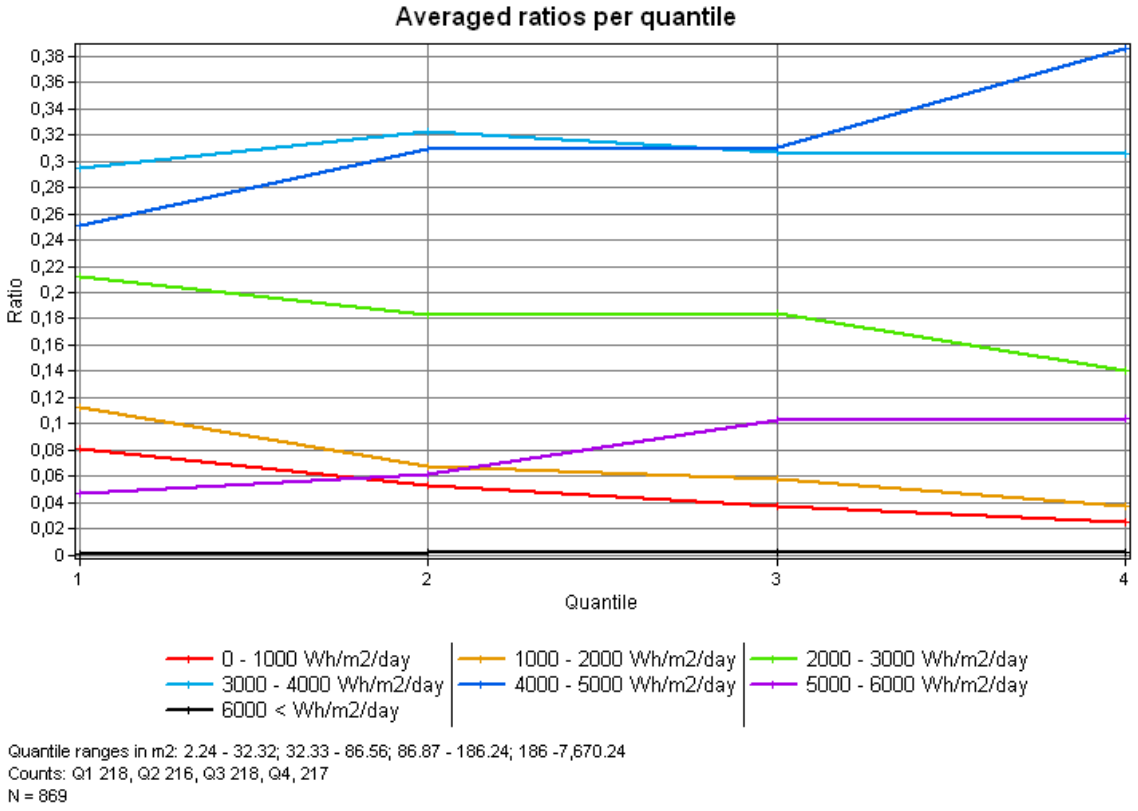
Boxplots: 0 - Sample 6051, 1 - Sample 8154, 2 - Sample 8279, 3 - Sample 23801, 4 - Sample 24359
Counts: 0 - 225, 1 - 238, 2 - 265, 3 - 36, 4 - 105

There are outliers present across the ranges, represented by the circles and stars. Outliers are not surprising, it reflects the fact that some roofs are in a worse context or have a less favorable roof structure compared to others and vice versa. For example, the outliers in the boxplots for the range 0-1000 have most of their roof likely covered in shadow, caused by neighboring objects such as trees or adjacent larger objects. This has a consequence that most radiation comes from diffuse radiation reducing the irradiation considerably. Outliers in the range 5000-6000 Wh/m2/day, could indicate a flat roof or slanted roof with a south facing orientation, unaffected by shadow. Thus, most roof area for these cases receive a high amount of direct radiation increasing the irradiation significantly.

The sampling areas represent low-density low-urban development, but we can conclude by examining the spread of the boxplots that the roof structures themselves are highly heterogeneous. Chart 6 (page 46), presents the distribution of factors through boxplots again, but now they are quantized based on roof size in m^2. Here a different trend can be discerned. Up to the range of 2000-3000, the boxplots of quantile 2 and 3, representing roofs above $86m^2$, have a smaller proportion of their area in these lower ranges, compared to quantile 0 and 1 representing roofs up to $86m^2$. For the range of 3000-4000 Wh/m2/day the boxplots are roughly equal, but from the 4000-5000 to the 5000-6000 Wh/m2/day range the 3rd and 4th quantile generally have higher ratios in these ranges compared to 1st and 2nd quantile.

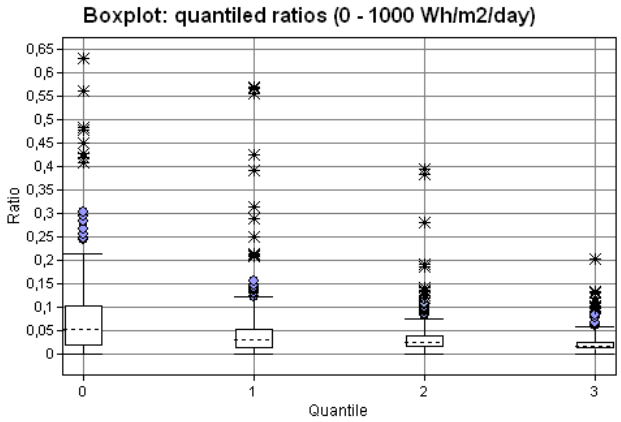
Graph 1 plots the averaged ratio per range of Wh/m2/day for each quantile and visualizes the trend more clearly. The lower quantiles have proportionally more area in the lower ranges of Wh/m2/day compared to the higher quantiles (see the red, orange and green lines). While higher quantiles have proportionally more area in higher ranges of Wh/m2/day compared to the lower quantiles (observe the dark blue and purple lines). This graph presents the homogenized averages for cluster 0 of the low-rise urban development class. In the section on validation these are tested by applying them to two test areas.

Graph 1 Averaged ratios per Wh/m2/day range plotted against quantiles

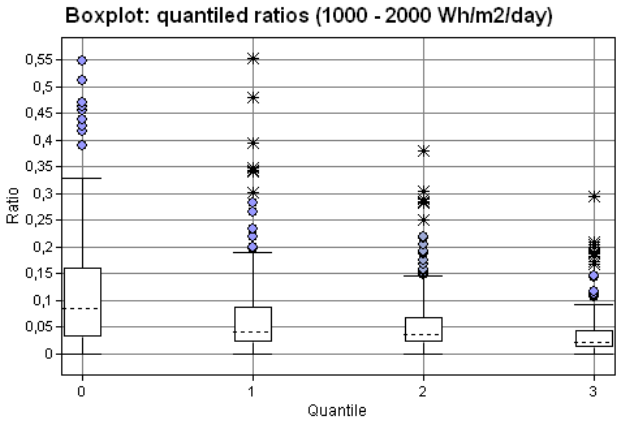


Quantile ranges in m2: 2.24 - 32.32; 32.33 - 86.56; 86.87 - 186.24; 186 -7,670.24
Counts: Q1 218, Q2 216, Q3 218, Q4, 217
N = 869

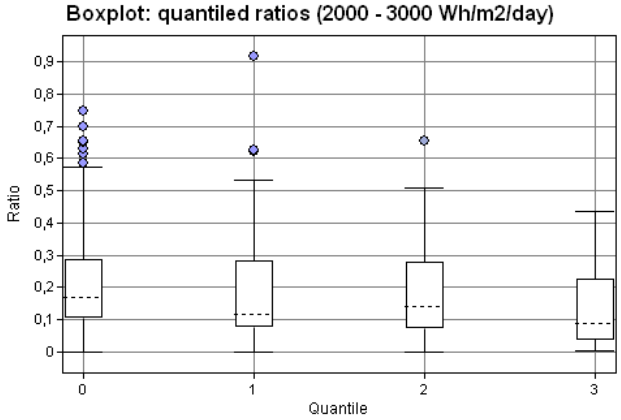
Chart 6 Distribution of ratios across quantiles sample areas



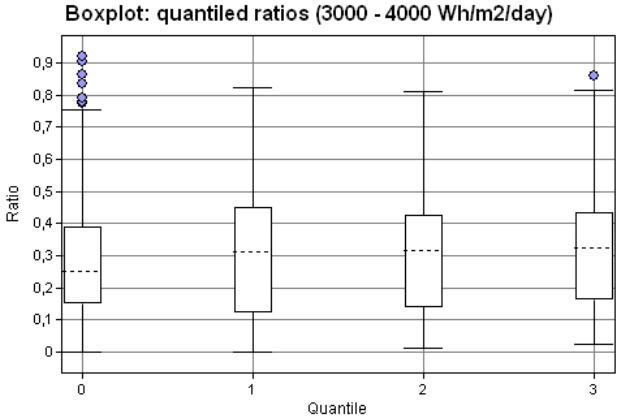
Quantile ranges in m2: 2.24 - 32.32; 32.33 - 86.56; 86.57 - 186.24; 186.24 - 7,670.24
Counts: Q1 217, Q2 216, Q3 218, Q4 216
N = 869



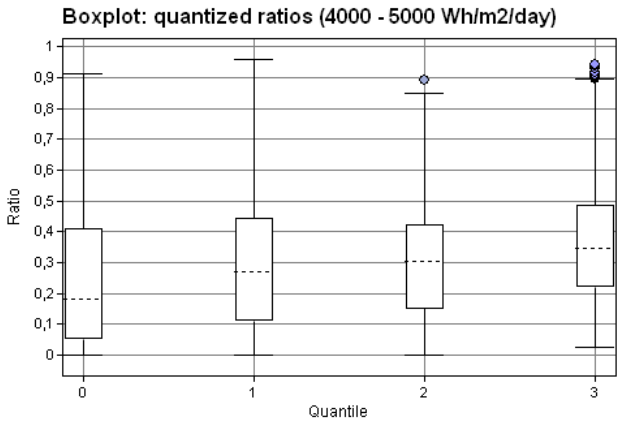
Quantile ranges in m2: 2.24 - 32.32; 32.33 - 86.56; 86.57 - 186.24; 186.24 - 7,670.24
Counts: Q1 217, Q2 216, Q3 218, Q4 216
N = 869



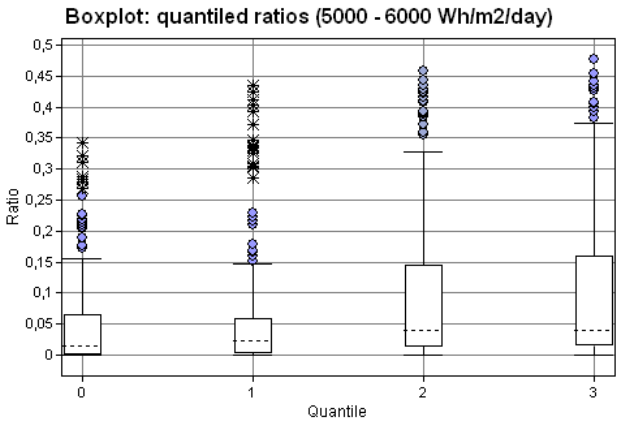
Quantile ranges in m2: 2.24 - 32.32; 32.33 - 86.56; 86.57 - 186.24; 186.24 - 7,670.24
Counts: Q1 217, Q2 216, Q3 218, Q4 216
N = 869



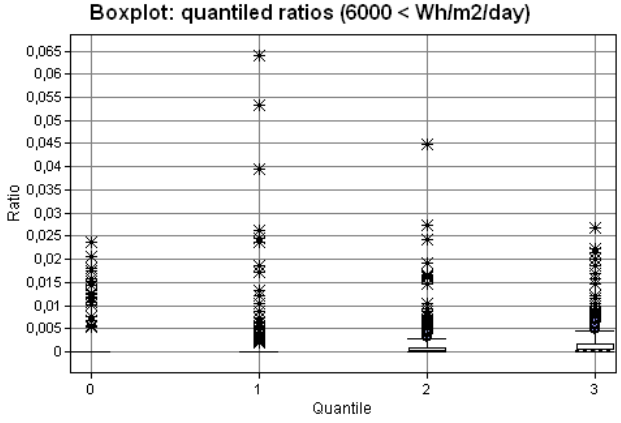
Quantile ranges in m2: 2.24 - 32.32; 32.33 - 86.56; 86.57 - 186.24; 186.24 - 7,670.24
Counts: Q1 217, Q2 216, Q3 218, Q4 216
N = 869



Quantile ranges in m2: 2.24 - 32.32; 32.33 - 86.56; 86.57 - 186.24; 186.24 - 7,670.24
Counts: Q1 217, Q2 216, Q3 218, Q4 216
N = 869



Quantile ranges in m2: 2.24 - 32.32; 32.33 - 86.56; 86.57 - 186.24; 186.24 - 7,670.24
Counts: Q1 217, Q2 216, Q3 218, Q4 216
N = 869



Quantile ranges in m2: 2.24 - 32.32; 32.33 - 86.56; 86.57 - 186.24; 186.24 - 7,670.24
Counts: Q1 217, Q2 216, Q3 218, Q4 216
N = 869

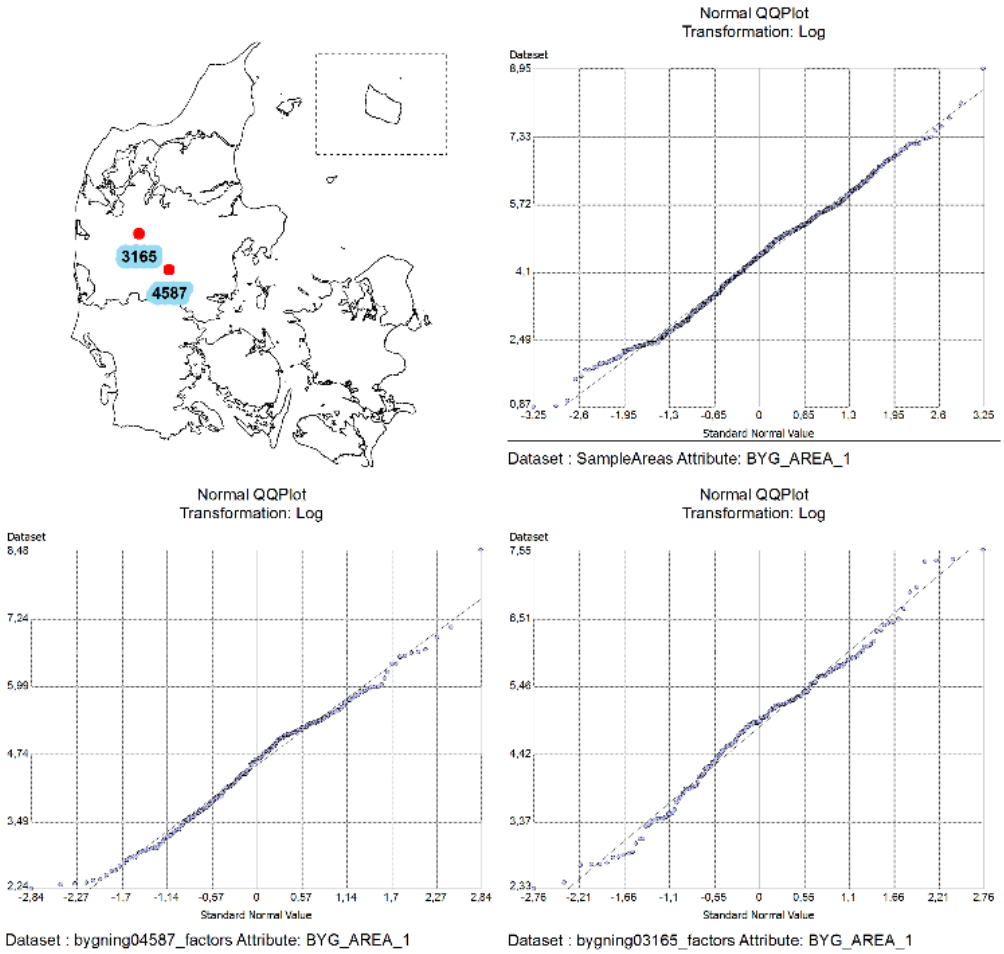
4.4 Validation of the ratios

In this section, the averaged ratios of the sampled areas, or training dataset, are tested. They are applied to two randomly picked test areas from the same cluster and class, cluster 0 of the low-rise urban development, and the results are compared against modelled results. Table 8 present some descriptive statistics on these test areas that includes the count of buildings and total, average, maximum, minimum and roof area. In chart 7 the distribution of roof area in m^2 of the sampled areas is compared with those of the test areas separately. In the latter graph, the locations of the test areas are likewise depicted. The normal QQ plots indicate that the distributions of roof area in m^2 of the test areas are similarly lognormal, although the highest and lowest values show slightly different deviations to the expected lognormal distribution, compared to the combined sampled areas.

Table 8 Descriptive statistics of tested samples

Sample ID	*Density*	*Count*	*Total roof area*	*Avg. roof area*	*Min. roof area*	*Max. roof area*
3165	0.00492	147	30,748	209	10	1,658
4587	0.00459	209	28,696	137	9	1,161

Chart 7 Distribution of roof area size in m2 per test sample and their location



The test areas have gone through the same solar radiation modelling process and calculation of the size in m^2 of each portion of roof, receiving a range of Wh/m2/day. The predicted roof portions are compared against the measured roof portions per range of Wh/m2/day. The predicted portions are based on an application of the averages presented in Graph 1 (page 44) and Table 9 below. Before they are applied, the buildings present in the test areas are grouped according to the same quantile ranges applied to the training dataset.

Table 9 Average ratios per quantile across the different Wh/m2/day ranges

Quantile	*Count*	*Avg ratio range 1*	*Avg ratio range 2*	*Avg ratio range 3*	*Avg ratio range 4*	*Avg ratio range 5*	*Avg ratio range 6*	*Avg ratio range 7*
1	218	0.0814	0.1124	0.2120	0.2947	0.2509	0.0476	0.0010
2	216	0.0537	0.0683	0.1830	0.3221	0.3096	0.0613	0.0019
3	218	0.0371	0.0577	0.1840	0.3063	0.3106	0.1025	0.0017
4	217	0.0251	0.0373	0.1340	0.3056	0.3856	0.1043	0.0021

For both test areas, table 10 contains, per range, the root-mean-square error (RMSE) and mean absolute error (MAE) between the measured and predicted portions in m^2 for each test area. As expected, the errors increase from the lower to the higher quantiles, but the errors are also larger when the middle ranges are concerned - from 2000-3000 Wh/m2/day up to 5000-6000 Wh/m2/day. This increase in the prediction error is explained by the boxplots. The latter show an increased variance for these ranges, which an average ratio cannot very well account for as can be seen here.

Table 10 RMSE and MAE values for both test areas

Test area ID: 4587

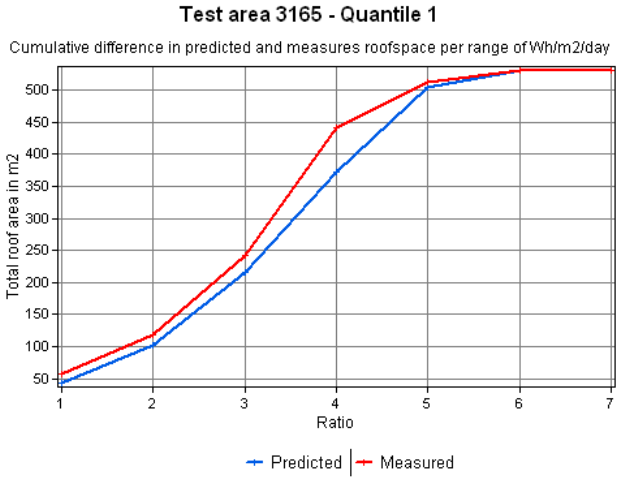
Quantile	**Test**	**Ratio 1**	**Ratio 2**	**Ratio 3**	**Ratio 4**	**Ratio 5**	**Ratio 6**	**Ratio 7**
1	RMSE	3.398	3.765	3.491	4.101	4.571	0.958	0.043
Count: 40	MAE	1.769	2.283	2.450	3.201	3.665	0.760	0.029
2	RMSE	4.201	6.312	8.671	11.728	11.963	6.189	2.154
Count: 64	MAE	2.121	3.906	6.823	9.029	10.458	3.795	0.403
3	RMSE	3.219	7.226	19.482	25.423	26.679	19.875	1.234
Count: 58	MAE	2.423	5.929	16.564	21.408	22.296	15.612	0.506
4	RMSE	12.864	27.622	82.694	62.179	126.064	60.686	11.019
Count: 58	MAE	6.991	15.639	53.411	44.825	78.494	40.155	3.019

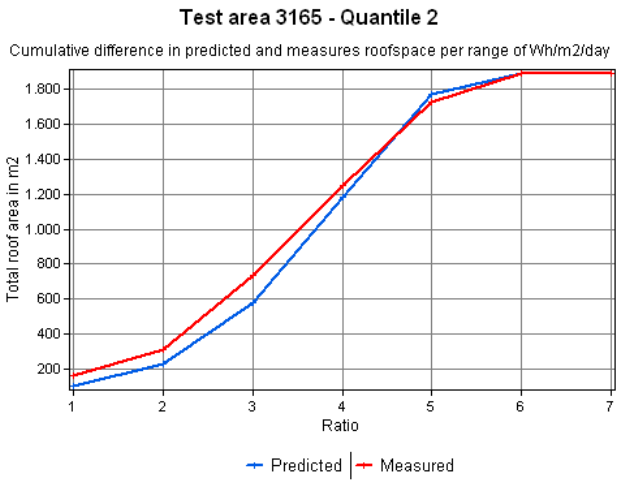
Test area ID: 3165

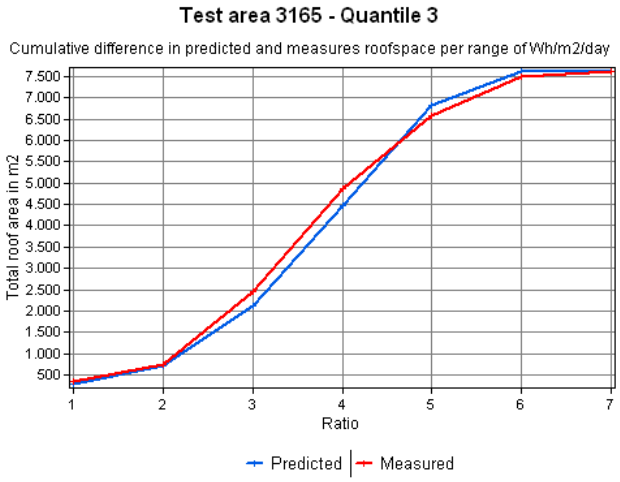
Quantile	**Test**	**Ratio 1**	**Ratio 2**	**Ratio 3**	**Ratio 4**	**Ratio 5**	**Ratio 6**	**Ratio 7**
1	RMSE	1.643	2.454	3.418	5.145	3.869	1.180	0.063
Count: 24	MAE	1.110	1.651	2.231	3.987	3.320	0.964	0.034
2	RMSE	5.971	4.422	8.989	12.434	13.448	6.927	0.134
Count: 33	MAE	3.881	3.429	6.840	10.877	11.743	4.804	0.120
3	RMSE	6.518	8.435	22.109	26.557	27.168	19.343	8.333
Count: 56	MAE	3.049	5.608	17.605	22.390	23.493	14.822	2.094
4	RMSE	15.663	20.295	83.560	93.982	186.473	99.439	2.507
Count: 60	MAE	8.199	14.280	58.905	65.954	113.128	66.986	1.103

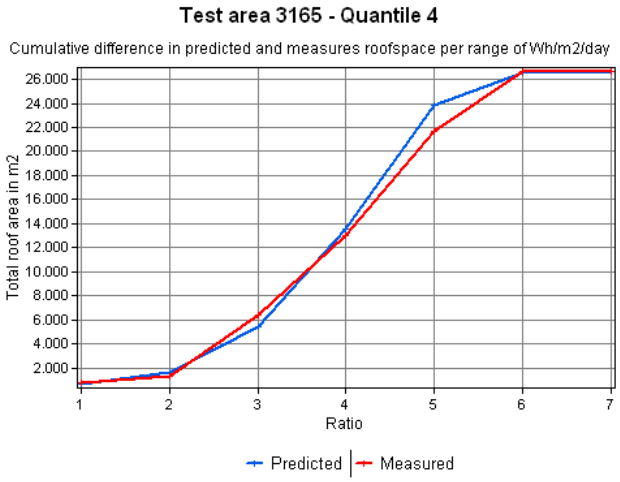
Below are a set of cumulative charts (chart 8), comparing the cumulative measured and predicted rooftop areas in m^2 across each ratio per quantile for both test areas. It puts the error margins into perspective, and the predicted trend is not significantly different from that what is measured, with a few exceptions. In test area 3165 the proportion of the lower ranges tends to be underestimated in quantile 1 and 2. In the chart for quantile 1, ratio 4 (3000-4000Wh/m2/day) this underestimation is the largest but amounts to a cumulative difference of roughly 75 meters. In quantile 4 the proportion of 4000-5000Wh/m2/day (ratio 5) is overestimated resulting in a difference of roughly 2,2000m^2, which is the largest error in absolute terms.

Chart 8 Predicted and measured cumulative roof portion's area in m2 for test area 3165









Below are the charts for test area 4587 (chart 9). Similar to test area 3165, the cumulative differences are not very large. The lower ranges tend to be underestimated in quantile 1 and 2, resulting in a lower cumulative area until the 5th ratio. In quantile 3 and 4 we see that the 5th ratio is likewise here overestimated.

Chart 9 Predicted and measured cumulative roof portion's area in m2 for test area 4587

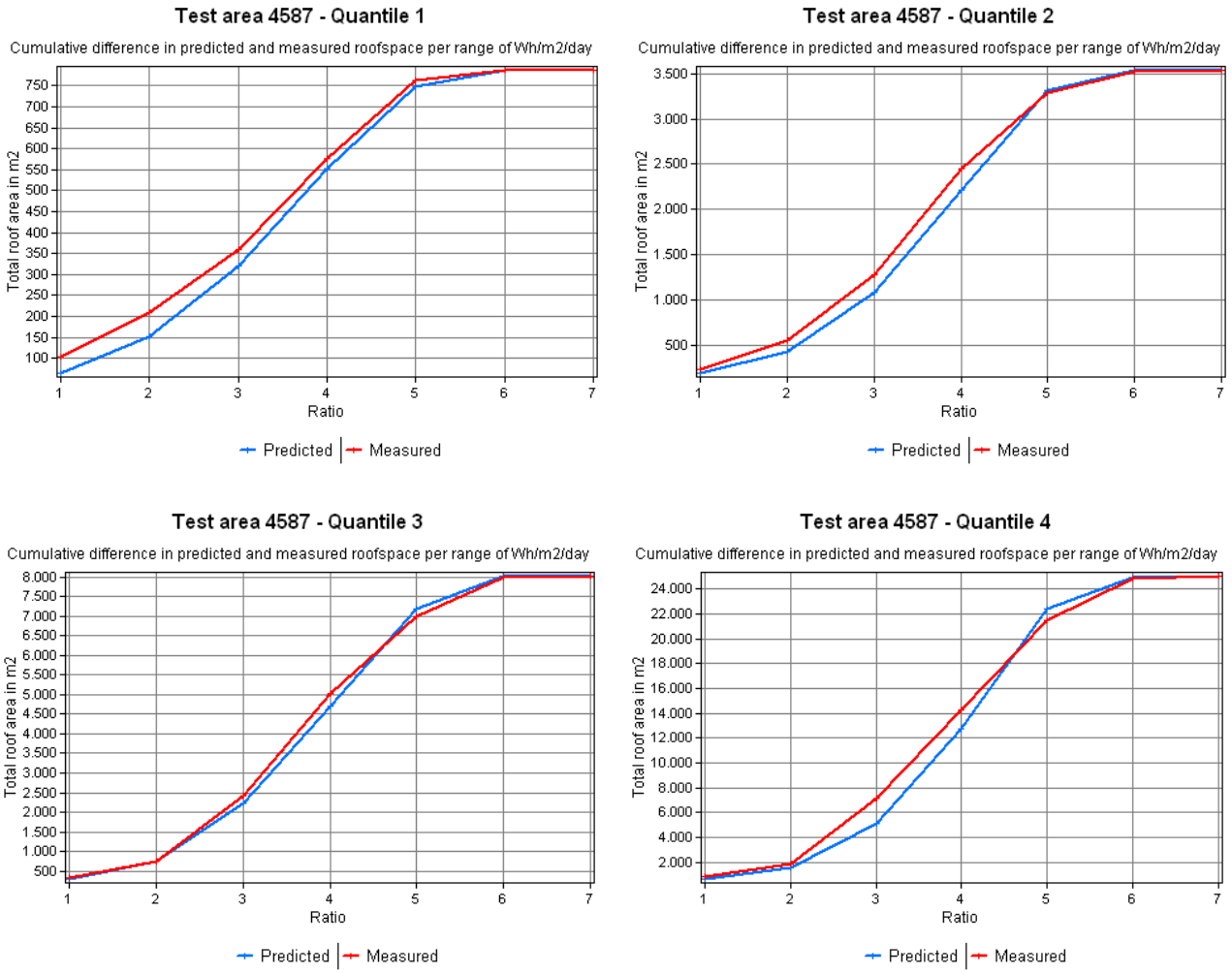


Table 11 and table 12, contain the measured and predicted total and average roof area in m^2 across the different ratios for each test area, in absolute terms. In both tables those cells where the prediction overestimated the area, are highlighted in light red and where it underestimated the area, are highlighted in green. In absolute terms Q3 and Q4 show the largest differences and the predictor is of by several hundred up to more than a thousand square meters, starting from Ratio 3 to Ratio 6.

Table 11 Absolute differences between measured and predicted values per Wh/m2Day range (test area 4587)

	Test ID: 4587	*Ratio 1*	*Ratio 2*	*Ratio 3*	*Ratio 4*	*Ratio 5*	*Ratio 6*	*Ratio 7*
Q 1	Total measured	102.72	106.72	150.24	215.52	188	24	0.48
	Avg. measured	2.57	2.67	3.76	5.39	4.7	0.6	0.01
	Total predicted	64.08	88.50	167.02	232.11	197.65	37.50	0.81
	Avg. predicted	1.6	2.21	4.18	5.80	4.94	0.94	0.02
Q 2	Total measured	233.60	318.56	726.24	1,166,90	847,36	224,96	20.00
	Avg. measured	3.65	4.98	11.35	18.23	13.24	3.52	0.31
	Total predicted	190.07	241.64	647.34	1,139.33	1,095.35	216.95	6.92
	Avg. predicted	2.97	3.78	10.11	17.80	17.11	3.39	0.11
Q 3	Total measured	329.60	422.56	1,663.36	2,618.56	1,965.92	996	23.04
	Avg. measured	5.68	7.29	28.68	45.15	33.90	17.17	0.40
	Total predicted	297.83	462.43	1,475.88	2,456.03	2,490.84	822.05	13.98
	Avg. predicted	5.13	7.97	25.45	42.35	42.95	14.17	0.24
Q 4	Total measured	797.44	1,036.32	5,321.44	7,141.92	7,256	3,361.28	154.72
	Avg. measured	13.75	17.87	91.75	123.14	125.10	57.95	2.67
	Total predicted	628.76	935.20	3,508.55	7,660.82	9,666.93	2,615.91	52.92
	Avg. predicted	10.84	16.12	60.49	132.08	166.67	45.10	0.91

For the second test area, the overestimated and underestimated ratios are quite similar, reflecting the cumulative graphs from before. The higher quantiles are here most affected as well. In particular Q4, where Ratio 3 and 6 are underestimated by 1,300m^2 and roughly 2,200m^2 respectively, and Ratio 4 and 5 are overestimated by roughly 1,500m^2 and 1,650m^2 respectively.

Table 12 Absolute differences between measured and predicted values per Wh/m2Day range (test area 3165)

	Test ID: 3165	*Ratio 1*	*Ratio 2*	*Ratio 3*	*Ratio 4*	*Ratio 5*	*Ratio 6*	*Ratio 7*
Q1	Total measured	56.48	62.08	122.88	199.36	71.36	18.08	0.32
	Avg. measured	2.35	2.59	5.12	8.31	2.97	0.75	0.01
	Total predicted	43.16	59.61	112.50	156.35	133.13	25.26	0.55
	Avg. predicted	1.80	2.48	4.69	6.51	5.55	1.05	0.02
Q2	Total measured	164.64	148.00	426.08	512.80	474.88	164	1.44
	Avg. measured	4.99	4.48	12.91	15.54	14.39	4.97	0.044
	Total predicted	101.65	129.22	346.19	609.29	585.77	116.02	3.70
	Avg. predicted	3.08	3.92	10.49	18.46	17.75	3.52	0.11
Q3	Total measured	344.64	394.72	1,714.72	2,406.08	1,717.92	933.28	112.00
	Avg. measured	6.15	7.05	30.62	42.97	30.68	16.67	2.00
	Total predicted	283.13	439.62	1,403.06	2,334.84	2,367.94	781.49	13.29
	Avg. predicted	5.06	7.85	25.05	41.70	42.28	13.96	0.24
Q4	Total measured	750.40	617.12	5,039.84	6,633.76	8,605.92	4,957.12	51.04
	Avg. measured	12.51	10.29	84	110.56	143.43	82.62	0.85
	Total predicted	668.54	994.37	3,730.53	8,145.51	10,278.54	2,781.42	56.27
	Avg. predicted	11.14	16.57	62.18	135.76	171.31	46.36	0.94

To conclude this chapter. The stratification process has reduced the within class heterogeneity, in terms of roof area density, into relatively homogeneous groups of potential sampling areas. A closer examination of the buildings contained in the random sample taken from cluster 0 of the low-rise urban development class, has shown that in terms of roof area in m^2, they all have a lognormal distribution. This means that besides density, it can be expected that the sampling areas themselves have a similar diversity of buildings in terms of roof area. However, the clear-sky global solar radiation modelling results, indicate that this stratification cannot completely account for the heterogeneity of the roof structures and immediate surroundings, and the effect they have on the rooftop solar energy potential. This was first shown by the spread of the ratios presented in the boxplots in chart 4 and 5 (page 43 and 44). These indicated that there is a high variance which, when quantized based on roof area remains.

When the ratios per rooftop part of the sampled areas are averaged across quantiles (Graph 1, page 45), a trend emerges in which larger buildings have a higher potential in terms of solar energy compared to smaller buildings. In the validation section this trend is tested, but the application of the averaged ratios in the two test areas result some significant RMSE and MAE values, particularly in those ranges that have a high variance in their contributions to the total roof area (3000-4000, 4000-5000 and 5000-6000 Wh/m2/day). When comparing the differences in absolute terms, the cumulative charts highlight where the main differences are, but overall the predicted curve does not deviate to much from what is measured. Examining the differences between the ratio's themselves the differences become more pronounced, especially for the 3000-4000 and 4000-5000Wh/m2/day range, where the predicted contribution to the total area can be off up to 2.2km^2.

5. DISCUSSION

5.1 Data suitability

The main data inputs used are a DSM, 2D building footprints that represent the roof outlines, urban development and land-use polygons. The high-resolution DSM is suitable in modelling roof structures and neighboring objects. As input in the solar radiation modelling, it can provide detailed results of the impact of both the roof structure and neighboring objects on modelled clear-sky global solar irradiation. While objects smaller than 40 by 40 centimeters cannot be evaluated, their impact is likely limited on the solar energy potential of a roof. The DSM is a source of error due to the way it is recorded and processed: not all values are recorded perpendicular to the ground, meaning that some height values represent vertical surfaces. This is unavoidable but will introduce a small error as some pixels will have been reclassified and added to roof portions which are facades. The high-resolution DSM has an impact on the processing time and storage. It had as a result that the methodology could only be run for a small part of the total potential estimation – a single cluster of one class. The DSM could be resampled to a lower-resolution, which would not only reduce the processing time and required storage space, but would also reduce the accuracy of the model.

The dataset with building footprints, is the closest approximated data available on roof area in Denmark. However, since they represent building footprints, they are not in all cases synonymous with rooftops, thus introducing a slight error that has not been quantified in this thesis. Another method would be to determine roof area through image classification algorithms, but this approach was outside of the scope of this thesis to include. Besides approximating roof areas, the timeliness that the dataset is constructed compared to the DSM, means that there are possible discrepancies between the two. It may have been the case that some of the outliers found by calculating the ratios across the modelled sample areas were discrepancies between the DSM and the footprints but this was not assessed.

The urban development and land-use polygons offer a more accurate and diversified classification of the stock of buildings in Denmark, compared to CORINE. A more accurate classification could be achieved if it was possible to link each building to a specific type of owner or use. This information exists in a national *Bygnings-og Boligregistret* (BBR)[33]. This data consists of geolocated point data that contain extensive information on most buildings of Denmark. Only parts of the footprint dataset is linked to the BBR, not all available polygons. Perhaps in the future they will become more closely coupled, which will allow for an extremely accurate classification[34].

5.2 On the performance of the sampling strategy

A sample based approach was necessary due to practical limitations in terms of time, storage capacity and processing power. The aim of the sampling strategy was to operationalize the assumption applied in this

[33] *Bygnings-og Boligregistret*(BBR) webpage: http://bbr.dk/

[34] The building use types in the BBR are even more extensive than the classes used here, see for example: http://bbr.dk/byg-anvendelse/0/30.

thesis, which was introduced in the introduction: “Buildings of a comparable size and in a like context, will on average receive similar amounts of solar radiation in Wh/m2/day on similar proportions of rooftop area.” It is based on the assumption applied by Lehmann and Peters (2003) and Izquierdo (2008) who rely on the relationship between population and/or building density and (available) rooftop area.

The way in which this assumption is put into practice reduced the heterogeneity of the building stock. First by stratifying them into classes, then by quadrat sampling and further stratification based on roof area density before they were sampled. A closer inspection of the random samples show that the strategy can return representative samples of the cluster they were taken from. Both in terms of roof area density, through the application of the dependent and/or Wilcoxon signed rank test, and the distribution of roof area in each sample area, based on the shared lognormal distribution between the samples.

The quadrat sampling with the use of a 2.5km by 2.5km gridded surface, while not directly compared to the impact of other sizes, is able to adequately sample the point data and make the solar radiation modelling a more manageable process at the same time, considering data handling and processing time. However, in low-density areas the modelling process currently spends too much time modelling irrelevant space to far away from roof areas that likely have no impact on them. But it allows for a uniform way of sampling the point data and from there to create clear stratifications based on density. If the grid cells are to large this spatial variation may become too generalized and at the same time increase the processing time and storage required.

The solar radiation modelling results, show that while in terms of roof area the samples are relatively representative of each other, in terms of roof structure this is less the case. To reiterate, no information is provided with the building footprints that characterize it’s roof to consider in the sampling process. The high variance in the different ratios (see the boxplots in chart 4 and 5, page 43 and 44), show that with the applied sampling strategy there is still a degree of heterogeneity with respect to roof structure, even when local conditions are similar and buildings are of a comparative size (Chart 6, page 46). This is not surprising as it was assumed that the building stock is heterogenous, but indicates that the assumption applied by Lehmann and Peters (2003) and Izquierdo (2008) must be implemented with caution. The inclusion of an error estimation by Izquierdo (2008) is one way of providing insight into the degree of over or under estimation that can result from the application of this assumption. Here a validation method is included to likewise provide an insight into the quality of the solar energy potential estimation.

The high variance is a direct result of the effect that different orientations, slopes and height of a roof and neighboring objects can have on a roof’s solar energy potential. However, on average a trend can be discerned in the sampled area, which is also quite instinctive: smaller buildings receive on proportionally larger areas of their roof lower ranges of clear-sky global solar irradiation, compared to larger buildings. While larger buildings receive on proportionally larger areas of their roof higher ranges of clear-sky global solar irradiation, compared to smaller buildings. This trend was visualized in graph 1 on page 45.

Regarding validation, the impact of this unaccounted-for heterogeneity became clearer. As the variance increases, the averaged ratios per quantile are not able to fully account for the heterogeneity of the different roof structures and end up underestimating and overestimating the size of the roof portions and consequently the solar energy potential of a rooftop. The prediction errors also increase when larger buildings are concerned as shown by the RMSE and MAE values in table 10 (page 48). However, these are only the results of a single cluster from a single class and it will have to be tested on other stratums for the results to be conclusive.

If these early results are seen in aggregated terms, depicted by the cumulative charts shown on page x, the predicted values do not deviate too much from the measured values. The aim of the thesis was not to provide a dataset that would allow for a per building assessment of its own solar energy potential (like the Sunmapper project), but rather to do so at aggregated levels for renewable energy planning purposes. From this perspective, these results look promising and with the included validation puts the estimation into perspective as to what it may overestimate and underestimate at these aggregated levels.

In the end, data that includes information on relevant variables such roof area, slope, orientation, height and neighboring objects are crucial to provide a proper estimation of the solar energy potential of rooftops. Without such data available or without an attempt at approximating these variables and their impact, will result into an estimation that is less accurate and provide less insight for policy makers and other third-parties interested in it. In particular, in study areas where it can be assumed that the building stock is heterogeneous. However, data availability is dependent on the context a study is conducted in. Therefore, in such cases where data is limited approximated data must be used (like CORINE) or an averaged coefficient from multiple other studies will have to be applied. While in other contexts such as the one presented here, well approximated roof areas are already available for a large-scale potential estimation.

5.3 Reproducibility and consistency through scripting

Diagram 1 (page 27) describes the entire process that is used to estimate the solar energy potential based on the applied assumption and validate it. Almost all major steps have been scripted with the use of Python 2.7 and several additional libraries mentioned in table 5 (page 32). By automating the process, it is easier to apply the methodology consistently, adapt it where necessary and keep track of the adaptations and their effects. There is a learning curve, but with the use of Python and several extensions it is possible to script data preprocessing, data handling, processing and modelling steps which provide considerable gains in efficiency, over manually performing each step. For example, by parallelizing the horizon calculations and solar radiation modelling, processing times were reduced by a factor of 3 and more gains are likely to be had in data handling.

6. CONCLUSION

The aim of this thesis was to develop a rooftop solar energy potential estimation for the country of Denmark – particularly a dataset containing individual buildings attribute data on rooftop area in m^2, estimated to receive a range of solar irradiation in watt hours per square meter per day (Wh/m2/day). Due to practical limitations, particularly time, this aim was not completely achieved and focus shifted to providing a validation of early results instead. This proved fruitful in gaining a better understanding of the applied methodology and how it performs.

Conclusions with regards to the state of GIS based potential estimation for solar conversion technologies can be found in the section 'Concluding the literature review' (section: xxx page x). To reiterate, the methodologies vary widely based on data availability, data quality, size and context of the study area, and practical limitations such as: time, storage capacity and hardware. What can be highlighted is that many studies lack the inclusion of a validation of the estimation results. This is often due to a lack of available data to do so. In such cases it is recommended to provide an estimation of the potential error and a method to do so is provided in this thesis. Arguably, without ground truth data it will put any estimation into question, but if such data is not available, the second-best option is to approximate this error. This will help users better understand the potential estimations, sources of error and the impact that they have on the results.

This thesis used freely available data which were limited in the information they provide on the object under study: the building rooftops and their solar energy potential. With the known variables of location and building footprint size, which was used as an approximation of the size of the rooftop, it was possible to make an estimation of the solar energy potential and estimate the prediction errors. In the process, it became clear that these variables are not able to fully account for the heterogeneity of the existing roof structures found in the study area. Such variables as: slope, orientation and height of a roof are besides its location and size crucial in estimating a solar energy potential. If such variables can be considered in a sampling approach, this will improve both the representativeness of sampled areas and the accuracy of the potential estimation.

Data limitations are unavoidable realities in all potential estimations also in the field of renewable energy. However, if it is possible to provide an understanding of the impact of these limitations the results can still be valuable for energy planning and policy making. Besides data, there are limitations based on time and available resources and an estimation is always a consideration between the depth and detail of the applied method and the accuracy of the results. The more depth and detailed the methodology the higher the accuracy of the result but the more time it will take to perform with an added increase in use of resources. If a balance must be struck, then it can be concluded that sampling is still a sound methodology in estimating the geographical potential of rooftop solar conversion technologies on a large scale.

7. Bibliography

Ahm, P., & Vedde, J. (2011). Large scale PV plants – Also in Denmark. Pa Energy Ltd, SciCon - Silicon & PV consulting. Retrieved from: solcelleforening.dk/site/wp-content/.../Large-Scale-PV-Plants-Also-in-Denmark.pdf

Assouline, D., Mohajeri, N., Scartezzini, J.L. (2017). Quantifying rooftop photovoltaic solar energy potential: A machine learning approach. *Solar Energy*, 141, 278-296. doi: http://dx.doi.org/10.1016/j.solener.2016.11.045

Bergamasco, L., & Asinari, P. (2011a). Scalable methodology for the photovoltaic solar energy potential assessment based on available roof surface area: Application to Piedmont Region (Italy). *Solar Energy*, 85(5), 1041-1055. doi: http://dx.doi.org/10.1016/j.solener.2011.02.022

Bergamasco, L., & Asinari, P. (2011b). Scalable methodology for the photovoltaic solar energy potential assessment based on available roof surface area: Further improvements by ortho-image analysis and application to Turin (Italy). *Solar Energy*, 85(5), 2741-2756. doi: http://dx.doi.org/10.1016/j.solener.2011.02.022

Brito, M.C., Freitas, S., Guimarães, S., Catita, C., Redweik, P. (2017). The importance of facades for the solar PV potential of a Mediterranean city using LiDAR data. *Renewable Energy,* 111, 85-94. doi: http://dx.doi.org/10.1016/j.renene.2017.03.085

Danmarks Meteorologiske Institut (DMI). (2012). 2001 – 2010 Design reference year for Denmark (Teknisk Rapport 12-17). Retrieved from http://beta.dmi.dk/fileadmin/Rapporter/TR/tr12-17.pdf.

Dubayah, R., & Rich, P. M. (1995). Topographic solar radiation models for GIS. *Int. J. Geographical Information Systems*, 9(4), 405-419.

FOT Danmark (2014). Specifikation FOT 5.1 forenklet udgave. Retrieved from: http://www.geodanmark.dk/Materiale/files/fot5/dk-pdf-fot51-forenklet.pdf

Fu, P., & Rich, P. (1999). Design and implementation of the Solar Analyst: An ArcView extension for modeling solar radiation at landscape scales. *Proceedings of the Nineteenth Annual ESRI User Conference*.

Fu, P., & Rich, P. M. (2002). A geometric solar radiation model with applications in agriculture and forestry. *Computers and Electronics in Agriculture*, 37, 25-35. doi: http://dx.doi.org/10.1016/S0168-1699(02)00115-1

Garnier, B. J., & Ohmura, A. (1968). A method of calculating the direct shortwave radiation income of slopes. *Journalof Applied Meteorology*, 7, 796-800. Retrieved from: http://journals.ametsoc.org/doi/pdf/10.1175/1520-0450(1968)007%3C0796:AMOCTD%3E2.0.CO;2

Garnier, B.J., & Ohmura, A., (1970). The evaluation of surface variations in solar radiation income. *Solar Energy*, 13, 21-34. Doi: 10.1016/0038-092X(70)90004-6

Ghisi, E. (2006). Potential for potable water savings by using rainwater in the residential sector of Brazil.*Building and Environment*, 41(11), 1544–1550.

Google. (2017). Project Sunroof data explorer: a description of methodology and inputs. Retrieved from: https://static.googleusercontent.com/media/www.google.com/en//get/sunroof/assets/data-explorer-methodology.pdf.

Guindon, B., Zhang, Ying, & Dillabaugh, C. (2004). Landsat urban mapping based on a combined spectral-spatial methodology. *Remote Sensing of Environment*, 92(2), 218–232.

Hetrick, W. A., Rich, P. M. & Weiss, S. B. (1993). Modeling insolation on complex surfaces. *Thirteenth Annual ESRI User Conference*, 2, 447-458. Retrieved from: http://professorpaul.com/publications/hetrick_et_al_1993_esri.pdf

Hetrick, W. A., Rich, P. M., Barnes, F. J. & Weiss, S. B. (1993). GIS-based solar radiation flux models. *GIS, Photogrammetry, and Modelling*, 3, 132-143. Retrieved from: http://professorpaul.com/publications/hetrick_et_al_1993_esri.pdf

Hofierka, J., & Kanuk, J. (2009). Assessment of photovoltaic potential in urban areas using open-source solar radiation tools. *Renewable Energy*, 34, 2206-2214. doi: http://dx.doi.org/10.1016/j.renene.2009.02.021

Hofierka, J., & Zlocha, M. (2012). A new 3-D solar radiation model for 3-D city models. *Transactions in GIS*, 16(5), 681-690. doi: 10.1111/j.1467-9671.2012.01337.x

Hong, T., Lee, M., Koo, C., Jeong, J., &Kim, J. (2016). Development of a method for estimating the rooftop solar photovoltaic (PV) potential by analyzing the available rooftop area using hillshade analysis. *Applied Energy*. doi: http://dx.doi.org/10.1016/j.apenergy.2016.07.001

Hoogwijk, M. M. (2004). On the global and regional potential of renewable energy sources. *Doctoral Research*,*University of Utrecht*. Retrieved from: http://dspace.library.uu.nl/bitstream/handle/1874/782/full.pdf?sequence=1

Huld, T., Müller, R., Gambardella, A., (2012). A new solar radiation database for estimating PV performance in Europe and Africa. *Solar Energy*, 86, 1803-1815. doi: http://dx.doi.org/10.1016/j.solener.2012.03.006

Izquierdo, S., Rodrigues, S., Fueyo, N. (2008). A method for estimating the geographical distribution of the available roof surface area for large-scale photovoltaic energy-potential evaluations. *Solar Energy*, 82(10), 929-939. doi: http://dx.doi.org/10.1016/j.solener.2008.03.007

Jakubiec, J. A., & Reinhard, C. F. (2013). A method for predicting city-wide electricity gains from photovoltaic panels based on LiDAR and GIS data combined with hourly Daysim simulations. *Solar Energy*, 93. 127-143. doi: http://dx.doi.org/10.1016/j.solener.2013.03.022

Jenks, George F. 1967. "The Data Model Concept in Statistical Mapping", International Yearbook of Cartography 7: 186–190.

Kabir, M. H., Endlicher, W., Jägermeyr, J. (2000). Calculation of bright roof-tops for solar PV applications in Dhaka Megacity, Bangladesh. *Solar Energy*, 35(8), 1760-1764. doi: http://dx.doi.org/10.1016/j.renene.2009.11.016

Khan, J., & Arsalan, M. H. (2016). Estimation of rooftop solar photovoltaic potential using geo-spatial techniques: A perspective from planned neighborhood of Karachi –Pakistan. *Renewable Energy*, 90, 188-203. doi:http://dx.doi.org/10.1016/j.renene.2015.12.058

Kondratyev, K.Y. (1970) Solar radiation and solar activity. Quarterly Journal Royal Meteorological Soceity. 96, 509- 522.

Kumar, L., Skidmore, K., & Knowles, E. (1997). Modelling topographic variation in solar radiation in a GIS environment. *Int. J. Geographical Information Science*, 10(5), 475-497. doi:http://dx.doi.org/10.1080/136588197242266

Latif, Z. A., Zaki, N. A. M., & Salleh, S. A. (2012). GIS-based estimation of rooftop solar photovoltaic potential using LiDAR. *IEEE 8th International Colloquium on Signal Processing and its Applications*.

Lehmann, H., & Peter, S. (2003). Assessment of roof & façade potentials for solar use in Europe. *Institute for sustainable Solutions and Innovations (ISUSI)*.

Li, Y., Ding. D., Liu, C., & Wang, C. (2016). A pixel-based approach to estimation of solar energy potential on building roofs. *Energy and Buildings*, 129, 563-573. doi: http://dx.doi.org/10.1016/j.enbuild.2016.08.025

McIntyre, J. H. (2012). Community-scale assessment of rooftop-mounted solar energy potential with meteorological, atlas, and GIS data: a case study of Guelph, Ontario (Canada). *Energy*, *Sustainability and Society*, 2(23). doi:10.1186/2192-0567-2-23

Möller, B., Nielsen, S., & Sperling, K. (2012). A Solar Atlas for Building-Integrated Photovoltaic Electricity Resource Assessment. Paper presented at 5th International Conference on Sustainable Energy and Environmental Protection, Dublin, Ireland.

Neteler, M. & Mitasova, H., 2008, *Open source GIS: A GRASS GIS Approach,* Third edition, Springer, New York, 406p.

Nunez, M. (1980). The calculation of solar and net radiation in mountainous terrain. *Journal of Biogeography*, 7(2), 173-186. doi: 10.2307/2844709

Ordonez, J., Jadraque, E., Alegre, J., & Martinez, G. (2010). Analysis ofthe photovoltaic solar energy capacity of residentialrooftops in Andalusia (Spain). *Renewable and Sustainable Energy Reviews*, 14, 2122-2130. doi:http://dx.doi.org/10.1016/j.rser.2010.01.001

Pillai, I. R., &Banerjee, R. (2007). Methodology for estimation of potential for solar water heating in a target area. *Solar Energy*, 81(2), 162–172.

Rich, P. M., Hetrick, W. A., Saving, S. C., & Dubayah, R. O. (1994). Using viewshed models to calculate intercepted solar radiation: Applications in ecology. *American Society for Photogrammetry and Remote Sensing Technical Papers.* 524-529. Retrieved from: http://professorpaul.com/publications/rich_et_al_1994_asprs.pdf

Robinson, N. (1966), *Solar Radiation.* Elsevier Publishing Company, Amsterdam.

Scartezzini, J.-L., Montavon, M., & Compagnon, R. (2002). Computer evaluation of the solar energy potential in an urban environment. *Paper presented at EuroSun 2002*. Bologna, Italy.
Ghosh, S., & Vale, R. (2006). Domestic energy sustainability of different urban residential patterns: A New Zealand approach. *International Journal of Sustainable Development*, 9(1), 16–37.

Scharmer, K., & Grief, J. (2000). *The European solar radiation atlas*. Les Presses de l'École des Mines. Paris. Retrieved from: http://catalog.mines-paristech.fr/Files/ESRA11res.pdf
Shreve,F. (1924). Soil temperature as influenced by altitude and slope exposure. Ecology. doi: 10.2307/1929010

de Smith, M. J., Goodchild, M. F., Longley, P. A. (2015) *Geospatial Analysis*. 5th ed. Online accessible via: http://www.spatialanalysisonline.com/HTML/index.html

Sun, Y., Hof, A., Wang, R., Liu, Jian, Lin, Y., Yang, D., (2013). GIS-based approach for potential analysis of solar PV generation at the regional scale: A case study of Fujian Province. *Energy Policy*. 58, p. 248-259. doi: http://dx.doi.org/10.1016/j.enpol.2013.03.002

Suri, M., & Hofierka, J. (2004). A new GIS-based solar radiation model and its application to photovoltaic assessments.*Transactions in GIS*, 8(2), 175-190. doi:10.1111/j.1467-9671.2004.00174.x

Suri, M., & Huld, T. A., & Dunlop, E. D. (2007). PV-GIS: a web-based solar radiation database for the calculation of PV potential in Europe. *International Journal of Sustainable Energy*, 24(2), 55-67. doi:10.1111/j.1467-9671.2004.00174.x

Suri, M., & Huld, T. A., Dunlop, E. D., &Ossenbrink, H. A. (2007). Potential of solar electricity generation in the European Union member states and candidate countries. *Solar Energy*, 81(10), 1295-1305. doi: http://dx.doi.org/10.1016/j.solener.2006.12.007

Wiginton, L. K., Nguyen, H. T., & Pearce, J. M. (2010). Quantifying rooftop solar photovoltaic potential for regional renewable energy policy. *Computers, Environment and Urban Systems*. 34(4), 345-357. doi:http://dx.doi.org/10.1016/j.compenvurbsys.2010.01.001

van Wijk, A. J. M. and J. P. Coelingh, 1993, Wind power potential in the OECD Countries, Utrecht University, Department of Science, Technology and Society, pp: 35. doi: 10.1002/qj.49705021008.

Williams, L. D., Barry, R. G., & Andrews, J. T. (1972). Application of computed global radiation for areas of high relief. *Journal of Applied Meteorology*. 11. Retrieved from: http://journals.ametsoc.org/doi/pdf/10.1175/1520-0450%281972%29011%3C0526%3AAOCGRF%3E2.0.CO%3B2

Ångstrøm,A.(1924). Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. doi: 10.1002/qj.49705021008.