D2.2 DATA GAPS AND QUALITY GAP ANALYSIS FOR FINANCIAL INSTITUTION DATA CHAINS

NEEM HUB

Nordic Energy Efficient Mortgage Hub

16 May 2023

PUBLICATION INFORMATION

SOURCE ACTIVITY: WP2, PUB
AUTHORS: P. VISETTI, J. AALTO, G. SINDERMANN, S. BERNT LINDEGAARD, S. VASILEVSKIS (GDFA, CENTER DENMARK)
FINAL DATE: 28.04.23
CONTRACTUAL DELIVERY DATE: MONTH 10

The Nordic Energy Efficient Mortgage Hub aims to scale-up lending to energy renovations in the Nordics and will publish a blueprint on how to accomplish this which will be implementable in other regions of Europe and, indeed, the world. In striving to increase energy renovations, the NEEM Hub will help achieve the targets of the European Green Deal and contribute to addressing ambitious national climate targets.

The NEEM Hub will be comprised of a long list of institutions from the financial sector, behavioural scientists, mortgage specialists and authorities, and digital technologies communities from across the Nordics, all guided by leading European economics consultancy, Copenhagen Economics. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 101032653.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

The success of the NEEM solution depends heavily on having high-quality data. This report assesses the data quality and potential data gaps and their effect on the NEEM solution.

The testing of the NEEM solution has been carried out successfully in the NEEM pilots. The model framework has been tested through verified in tests in Denmark, Norway and Sweden. Before considering scaling the solution, this report dives into the data itself, identifying the key data points, relevant data sources, and the most important aspects of this data. This includes an assessment of the quality of each data type in each Nordic country and its impact on the NEEM solution. Also, data gaps are considered.

Specifically, the categories of energy consumption data, static building data, and weather data are considered, and their implications are discussed. The data filtration process is also described as this key aspect includes only the relevant households that have all the required data available. Heating data is by far the most important bottleneck in Denmark and Norway, as central data hubs exist which can provide electricity, weather, and static building data of sufficient quality on a national or global level.

Heating data, however, has historically been less systematised in the district heating sector, making this the most fluctuating data point in the model. At the test sites, data gaps have been a challenge and adding new district heating data from other companies is expected to result in possible data gaps exceeding 10%, although there may be very few data gaps in some cases.

Additionally, the electrification of both heating and transportation requires additional work on the NEEM model as the impact of electric vehicles (EVs) and heat pumps can skew the results of the model. This is the primary threat to data quality across Nordic countries.

Chapter 1

INTRODUCTION

1.1 NEEM HUB AND NEEM CORE SOLUTION

As a part of the existing Energy Efficient Mortgages Initiative supported by Horizon 2020, the Nordic Energy Efficient Mortgage (NEEM) Hub aims to scale up lending for energy renovations in the Nordics. By increasing energy renovations, the NEEM Hub will help achieve the targets of the European Green Deal and contribute to addressing ambitious national climate targets.

To facilitate this, the project has developed a NEEM core solution: a three-step guide for banks to deliver customer-specific energy renovation recommendations to their customers. The solution is designed to simplify and automate the process of finding profitable energy renovations for residential houses and has been tested in multiple pilot studies in Denmark, Norway, and Sweden.

1.2 ABOUT THIS REPORT

1.2.1 Purpose and content

The success of the NEEM core solution depends heavily on having high-quality data. This report contemplates the data quality and potential data gaps and their effect on the solution. This deliverable covers technical reporting and offers a detailed explanation of the solution, with a focus on data foundation, availability, and quality. We identify the data needed to enable financial institutions to scale up green lending.

1.2.2 The data approach

To scale up the NEEM core solution for successful testing and deployment, creating a one-stop data access space for all the relevant data types was the highest priority. This report describes the current challenges and solutions related to the existing data foundation in the Nordic countries and the data availability in each country.

Chapter 2
NEEM CORE SOLUTION DATA FOUNDATION

Overall, there are three categories of data sources required to run the NEEM energy efficiency assessment models: energy consumption data, static building data, and weather data. This chapter offers a high-level overview of this underlying data, followed by chapters for each data type and country-specific details.

The NEEM core solution focuses on residential real estate limited to single-family, privately owned homes. To run the NEEM core solutions, it is crucial to combine the available consumption data with local weather data, data on the actual building area, and data on the number of building floors. It is also very important to know the details about the building's heating system and heating source type. Without these data points, it is not possible to estimate a building's energy performance accurately.

The main criteria specific to the NEEM demonstration sites for a household to be able to run the NEEM model is to ensure access to the right data. In Denmark, the project focuses on the district heating data from single-family houses located in the Triangle Region. In Sweden and Norway, the main acceptance criterion is also that the building is a single-family house. Otherwise, the NEEM model for these countries is adjusted to ingest the electricity data.

2.1 ENERGY CONSUMPTION DATA

The NEEM solution requires data on the energy consumption used for heating to estimate current energy efficiency. Heat consumption types may vary from building to building, typically being district heating or electricity. Moreover, there are large differences across the Nordics when it comes to households; in Denmark, for example, households rarely use electricity for heating. Other heating sources prevent the NEEM algorithm from predicting energy efficiency, as the consumption data are too infrequent and have too many uncertainties attached to them. In situations when the heating source of the consumption data is electricity, it is beneficial to know more about the heating system in the building itself, namely if it is regular electrical heating or a heat pump.

2.2 STATIC BUILDING DATA

Since the energy performance also depends on the building size, it is important to normalise the energy consumption per square metre to make a fair comparison. Necessary parameters comprise, among other things, the number of floors, floor size, total building area, location (latitude/longitude or postal code, address), and energy label. Further parameters could be relevant in the future, but the above-mentioned are sufficient to make a strong prediction.

The floor size, heat source, and location are the most essential parameters for the NEEM core solution. Additionally, the heating source plays an important role when determining the energy label based on the primary energy factor (PEF) coefficients. The heating source will determine the choice of values for PEFs in each country. The energy label is not crucial since it is what the NEEM core solution aims to estimate in a better and more data-driven way. Instead, the energy label is used as validation data to qualitatively compare how accurate the NEEM core solution is in estimating the energy performance when compared with the traditional energy-labelling methodology.

2.3 WEATHER DATA

The actual energy efficiency of a house depends on its performance when subjected to different weather characteristics such as wind, sunshine, and temperature. The NEEM algorithm primarily involves actual weather data, but, in some cases, the building's location needs additional parameters such as local terrain topology. One example is buildings in a very hilly or mountainous region, where wind and solar radiation may differ from house to house and area to area. This is especially relevant in the Norwegian case and in some areas of Sweden, whereas Denmark is very flat, and therefore the additional parameters are seldom required there.

2.4 MATCHING THE DATA SOURCES

A common identifier is required to create a dataset that matches specific building information data from public data sources with the energy consumption data from utility and weather data services. Utility companies often store specific clients' address strings in a slightly different format than in the publicly available building information database. This can cause some information loss when trying to match these address strings.

These concerns become important in situations when a specific utility company does not store building information data (i.e., area, floors) locally with the energy consumption data. Therefore, it is crucial to make a precise match with the public building information database. Regarding weather data, the geographical coordinates of buildings or, in some cases, postcodes (zip codes) are sufficient common identifiers to match data. This is not as critical as the previous step to match consumption data with static building data.

Chapter 3

ENERGY CONSUMPTION DATA

The NEEM core solution was developed and tested on two energy consumption data types – electricity and district heating. Each data source posed its own challenges. The data was retrieved from private utility companies, either directly or from centralised data hubs.

Most of the Nordic countries have a centralised IT system that gathers smart electricity meter data points in one place and thus facilitates data sharing via third-party access and user consent; Norway uses Elhub, and Denmark uses DataHub. Generally, these national data hub solutions require an electronic ID (eID) to use their API for data access. This was one of the limitations faced by the data preparation team since it was based in Denmark and could not utilise the Norwegian data hub to facilitate data ingestion. Therefore, the data preparation team had to make an individual agreement with the relevant electricity provider.

In the NEEM project, we had to establish several ad hoc agreements with utility companies to obtain the necessary data for our pilot test. As a consortium, we also explored how to establish agreements with national repositories, yet this proved to be more difficult. This was because many countries have legal requirements that restrict access to open-source national databases, necessitating the need for a registered legal entity within their borders to access the data. As a result, we had to navigate the legal requirements of each country to obtain the necessary permissions and establish the necessary agreements with third-party entities to acquire the data we needed. Although this process was time-consuming and complex, it was critical to the success of the NEEM project and enabled us to complete our pilot test successfully.

On the other hand, the district heating data, which is the most common building heating option in Denmark, did not have a centralised data hub. This means that to use district heating data for the NEEM core solution, it was necessary to have a separate agreement with each district heating company. Not having a central energy consumption data hub is one of the biggest hurdles to scaling the NEEM core solution for other countries.

To accommodate the more dynamic nature of the energy consumption of a building, it is important to have data with more frequent time steps. However, in the modelling process, the Technological University of Denmark (DTU) algorithm for NEEM core solutions was also adapted to predict the energy performance based on more coarse time series data, namely, the daily average consumption data. Another very important aspect of consumption data is the observation period, more specifically, whether the data covers the heating season of the year or not, and if it is at least half a year long. Of course, the greater the number of historical observations (multi-year time series data) each building with the same owner will have, the better the DTU algorithm's prediction will be.

3.1 DENMARK

Here, the NEEM core solution was tested on single-family buildings that have district heating as the main space heating solution. More than 1.8 million households, corresponding to two-thirds of all Danish homes, are already connected to district heating. The greatest challenge for this consumption data type is that there are approximately 400 district heating companies that are spread over the country. Around 50 of the companies are municipal utilities, which account for 50% of the total Danish district heating supply. The other half of the district heating is supplied by around 350 cooperatives. In addition, there are quite a few private companies.

3.1.1 Data availability

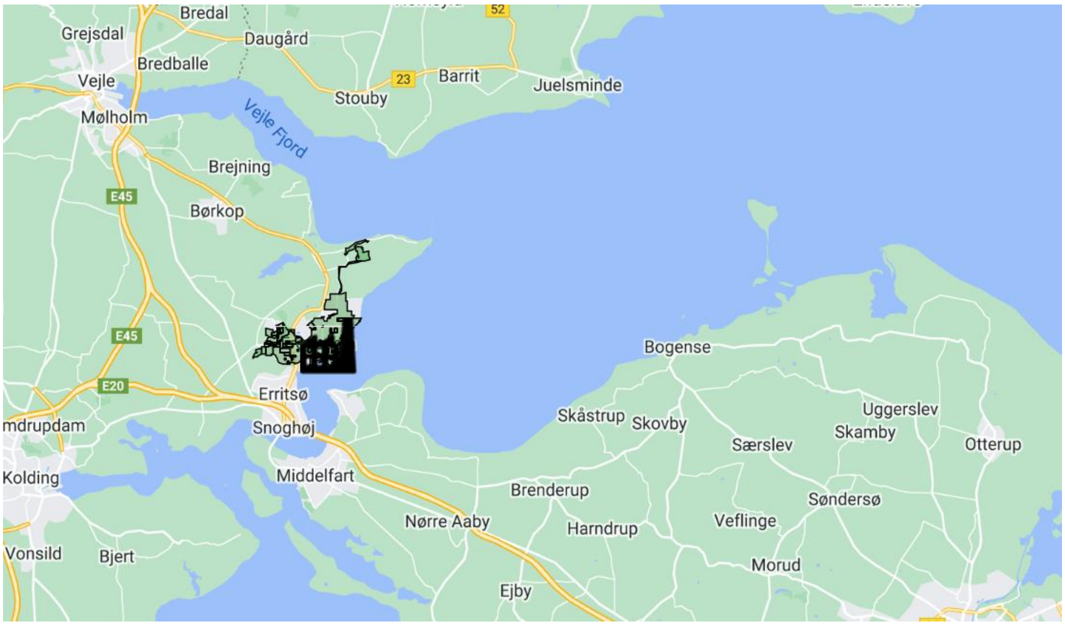
In general, the district heating providers in Denmark do not have a centrally accessible data-sharing solution like the one that is available for the electricity data. Individual clients can access their electricity data via a centralised API, Eloverblik, provided by Energinet. Eloverblik is a platform that is available to private individuals, businesses, and third parties. The purpose of the platform is to provide an overview of the data on electricity consumption and generation across Denmark and across electricity suppliers.

Having a district heating data hub would eliminate the extra step required to create individual data ingestion agreements with a large number of utility providers. Unfortunately, it is not yet possible in Denmark. Thus, one can say that the bottleneck of this demonstration site is the availability of the district heating data.

3.1.2 Data access, processing, and challenges

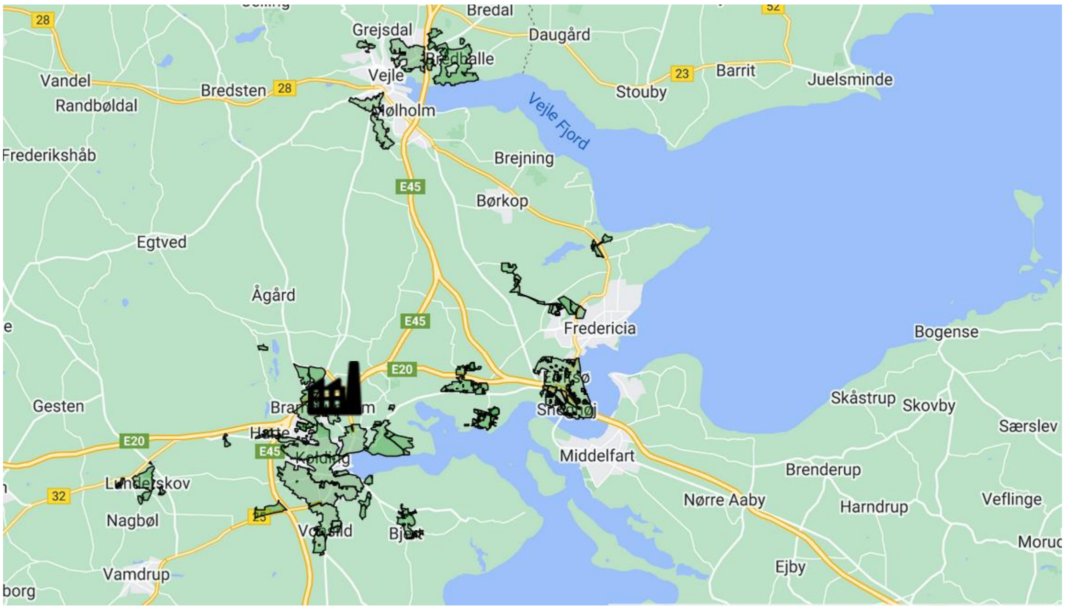
The NEEM core solution in Denmark was tested in buildings in the Triangle Region, mostly in the Fredericia municipality. Additionally, the NEEM project geographic coverage area in Denmark was outlined by the supply areas of the TREFOR Varme and Fredericia Fjernvarme utility companies. Both companies are the main district heating providers in the Triangle Region. The figures below show the aforementioned utility company supply areas.

Figure 1
Fredericia Fjernvarme coverage area



Source: DinGeo

Figure 2
TREFOR Varme coverage area



Source: DinGeo

Initially, the NEEM core solution was built on hourly district heating data, where Fredericia Fjernvarme data was the main data source in Fredericia municipality. To give an idea of the coverage of the Danish NEEM pilot, more than 25 000 households from the coverage area of Fredericia Fjernvarme and TREFOR Varme with district heating as a space heating solution could be evaluated with the NEEM core solution.

Since data with hourly resolution from TREFOR Varme was not fully ingested during the project, some of the households in the Fredericia area were not matched for the NEEM pilot case. It is also important to note that many of the district heating meters in TREFOR Varme used daily time series resolution. Therefore, the NEEM building energy efficiency prediction model had to be adjusted to accommodate time series data with the daily resolution. This is an example of a structural data quality issue – meaning that the DTU energy performance prediction model must be flexible enough to accommodate such differences in time series resolution.

3.1.3 Data quality

Providing district heating data for the DTU algorithm came with its own set of challenges. After providing the initial test dataset to DTU from Fredericia Fjernvarme with approximately 5500 single-family buildings, it was necessary to correct the data ingestion pipeline from TREFOR Varme to expand the demonstration pilot area since it was discovered that large amounts of older district heating meters with daily time resolution were not ingested in the Center Denmark data platform. This created some delays due to the need to organise extra meetings with the TREFOR Varme data specialists.

After solving the data ingestion problems, the team focused on data format and quality. TREFOR Varme and Fredericia Fjernvarme had varying layouts and formats that necessitated the implementation of individual data-cleaning pipelines. For each data source, it was necessary to place the relevant consumption data in separate files and rename the columns to match the input layout of the NEEM core solution requirements.

Additionally, some data-related challenges occurred when the building owners changed their utility company and the data from the new company were not yet available in the data lake, or were available too briefly. This situation may prevent the NEEM model from converging and may result in unreliable energy efficiency predictions due to a lack of data.

3.2 NORWAY

The NEEM core solution was tested on single-family buildings that have electricity as the main space heating solution.

3.2.1 Data availability

In Norway, an advanced metering infrastructure (AMI) for hourly meter readings was established for all households by the end of 2019. The electricity consumption and production data can be accessed through a data hub called Elhub. Elhub is a central IT system that supports and streamlines market processes such as electricity sales, move-in/-out, termination of supply, and the distribution and aggregation of metering values for all consumption and production in Norway. One can say that the Elhub data hub streamlines electricity data processing and market communications and it serves as a link between the smart meters and the customer.

3.2.2 Data access, processing, and challenges

In the case of Norway, the NEEM Hub wanted to leverage the data capacity of Elhub, which is the Norwegian data hub for the energy sector. This centralised system collects, stores, and manages data related to the production, consumption, and distribution of energy in Norway. Elhub was established to improve the efficiency and effectiveness of the energy market in Norway by enabling market participants to access accurate and timely data in a standardised and secure way. Elhub serves as a central platform for data exchange between energy companies, grid operators, and other stakeholders in the energy sector. It is owned and operated by the Norwegian IT company, Energi Data Service AS, and is regulated by the Norwegian Water Resources and Energy Directorate (NVE).

During the NEEM project, we recognised the importance of gaining access to Elhub's data, and, as such, we had several virtual meetings with the entity. However, Elhub required the organisation to be legally registered in Norway to gain access to its data. After exploring the possibility of registering in Norway, we managed to get Nordea Norway to establish an API gateway to Elhub. Nevertheless, even with the gateway in place, we encountered IT barriers and data privacy concerns when attempting to process the data in Denmark. Moreover, to retrieve the meter ID data, we needed Elhub's consent to gain access to all the meter IDs. The NEEM test volunteers would have had to approve this directly via the Elhub portal, and they would often be unaware of the existence of Elhub itself. Thus, test volunteers would have had to create a profile, register and log into the Elhub portal to allow us to gain access to their data. Doing this for every single household in the test seemed unrealistic and very ineffective. After extensive efforts, we had to abandon this avenue for testing due to these challenges. Moving forward, we believe future endeavours in Norway should focus on streamlining the process. We understand that SINTEF, the research institute, is exploring the creation of a digital environment with a direct link to Elhub to make the process easier.

3.2.3 Data quality

After it was decided not to pursue the data download option from Elhub using their API and third-party solution, access to the relevant consumption data was negotiated with Elvia, a Norwegian utility company. The data that was received from Elvia included only the buildings that were selected for the Norwegian pilot.

In general, the electricity consumption data from Elvia was of good quality. It was not necessary to join the metering data with publicly available static building data since the metering files already contained building area values. Concerning data formatting, it was necessary to place the consumption data in separate files and rename the columns to match the input layout of the NEEM core solution requirements.

On the other hand, since the consumption data did not contain precise addresses or coordinates, the weather data had to be retrieved using the municipality and postal codes. This data process is described in detail under the weather data description.

3.3 SWEDEN

During our data collection process in Sweden, we encountered challenges in obtaining real data agreements from utility companies. We reached out to multiple companies, including Vattenfall and E.ON, through various channels such as phone, email, website, and social media, but, in the end, we did not succeed in arranging a third-party agreement.

Moreover, an interesting endeavour was carried out with Sollentuna Energi & Miljö (SEOM) in Sweden. Sollentuna Energi & Miljö is a local energy and environmental company located in the municipality of Sollentuna, Sweden. The company provides electricity, heating, cooling, and waste management services to both private households and businesses in the area. Sollentuna Energi & Miljö is owned by the municipality of Sollentuna and aims to promote sustainable and environmentally friendly energy practices. In recent years, the company has invested in renewable energy sources such as wind and solar power, as well as in energy-efficient technologies to reduce its environmental impact. The company also offers advice and services to help its customers reduce their energy consumption and carbon footprint. Sollentuna Energi & Miljö offers its customers the possibility to download their energy consumption data from its website. This feature allows customers to transparently and easily keep track of their energy usage and make more informed decisions about their energy consumption. Therefore, the NEEM Hub used this feature to carry out a NEEM core solution pilot test with bank employees from the area.

In a pre-pilot test, we were able to obtain anonymised district heating data from Stockholm Exergi in the spring. The data primarily covers Stockholm, which has a high concentration of apartments and limited single-family houses, making it sufficient to run a pilot test of our model.

STATIC BUILDING DATA

In most cases, data related to the buildings is obtained publicly from centralised state-owned data sources. This data can generally be accessed using open-source API services or is directly downloadable. Depending on the implementation, it may be necessary to use a digital signature to gain access, which may complicate the process. For example, data about all buildings in Denmark are public and can be viewed and downloaded by anyone from Bygnings-og Boligregistret (BBR).

The utility companies are also able to provide some information about buildings but it will depend on each company's data storage system. Here, the utility companies may use the data from public data registries, therefore it may be outdated.

The overall expectation is that the data quality is very good since the information about properties is regulated and controlled by state-owned organisations. This expectation was met while working with the data from the Nordic building and property registries. One must be aware that sometimes the building data may be outdated due to the slow registration process. This is especially important when the building has been renovated and the heating source type changed, as this affects the DTU's algorithm prediction accuracy and NEEM core solution results.

3.4 DENMARK

3.4.1 Data availability

Data about all buildings in Denmark are public. This data can be viewed or downloaded from the Danish building registry, Bygnings-og Boligregistret (BBR).

As an alternative data source on building floor size, the number of floors, and information about the building's heating system, the energy performance certificate (EPC) report can be used. The EPC report contains all the relevant information about a building and is obtained via a certification process performed by an energy consultant from a certified company. One can search for information on the EPC of a particular building and download the EPC report from the Danish Business Authority (Ehrvervsstyrelsen) Boligejer website.

An energy performance certificate is mandatory and is required for the sale and rental of buildings, new buildings, and public buildings over 250 m^2. An EPC is valid for 10 years unless significant changes are made to the building, which will affect the energy performance. Therefore, it is important to check how recently the EPC report has been updated. One must be aware that a large amount of the energy label data in Denmark is missing.

It is important to mention that the energy label value itself is not directly used as input into DTU's algorithm and therefore is optional for the NEEM core solution if the static building data has already been obtained from the public registry.

3.4.2 Data access, processing, and challenges

The static building data from BBR was available for download using the public API, Danmarks Adressers Web API (DAWA). Here, all the relevant information was located in multiple tables. Therefore, it was necessary to download all the tables separately and perform multiple table joins to retrieve all the necessary data. Also, the EPC data of all Danish buildings was downloaded using Energistyrelsen's EMOData-service API. Even though energy label data is optional, this data was used by the DTU data scientists to verify their model predictions and gauge how far their energy label predictions were from the EPC results.

3.4.3 Data quality

On many occasions, the building registry data can be outdated. The responsible agency, the Danish Property Assessment Agency, is aware of this problem and is working on improving data quality.

Among other things, it was necessary to combine the address string from the BBR tables with the address strings from the consumption data table. To increase the address string matching ratio, it was necessary to normalise the strings. The process of address normalisation required some extra verification since the internal systems among the utility companies vary and, more often than not, do not meet the BBR standards. This increases the risk of error, especially when joining across data tables from different data sources.

3.5 NORWAY

3.5.1 Data availability

The Norwegian Mapping Authority (Kartverket) manages detailed public geographical information about Norway and also information about property registry data distribution for users and stakeholders. Here, one can search for information about addresses, buildings, and properties.

3.5.2 Data access, processing, and challenges

The limitation to accessing Kartverket building data is that one must have an electronic Norwegian ID solution to log in and complete the search. The same limitation applies to the energy label data. For a scale-up, this information is crucial, especially when considering the heat source data, which is crucial for the NEEM core solution. Another alternative is to retrieve this information from the utility company, if it is available.

3.5.3 Data quality

It was not necessary to use the Kartverket service since the electricity consumption data included the necessary building area values; no additional data processing was necessary for this data type. The values from the consumption data could be used directly in the DTU algorithm.

3.6 SWEDEN

The Swedish National Board of Housing, Building, and Planning (Boverket) is a central government authority that works with issues on how to plan society, buildings, and housing. The Swedish Mapping, Cadastral and Land Registration Authority (Lantmäteriet) manages all information on properties in Sweden. The Lantmäteriet registers contain up-to-date information on all properties in Sweden.

In principle, the building cadastre data is publicly available for access and it can be accessed in the following ways:

1. The Lantmäteriet website: The Swedish Mapping, Cadastral and Land Registration Authority's website provides access to the building cadastre data. One can search for the information on the website or download the data directly.
2. The Swedish open data portal: The building cadastre data is also available on the Swedish open data portal, Öppna Data. One can access the data and download it from this portal as well.
3. Through a third-party service provider: Several third-party service providers offer access to building cadastre data. These providers may offer additional features, such as data analysis tools, and may require a subscription or payment to access the data.

Nevertheless, to access building data in Sweden, such as floor size, number of rooms, and heating type, one typically needs to have a Swedish personal identity number (personnummer) or a Swedish organisation number (organisationsnummer). This is because building data is classified as personal information and is protected under the Swedish Personal Data Act (Personuppgiftslagen).

Individuals can access building data on their own properties using their personal identity number. It is important to note that accessing and using building data without proper authorisation or consent may be a violation of Swedish data protection laws and can result have legal consequences. Thus, this extra layer became a significant barrier in the Swedish NEEM core solution testing because it required individuals to download this information and give the consortium consent to utilise them.

3.6.1 Data access, processing, and challenges

As in Norway, to use the aforementioned e-services to extract building information, it was necessary to log in with national digital signatures or eID. This option was not investigated further due to the NEEM project boundaries and the limited demonstration case.

3.6.2 Data quality

The electricity consumption data included the necessary building area values, the number of floors, and heating system type descriptions. Therefore, no additional work was necessary for this data type and the values could be directly used in the DTU's energy performance prediction algorithm. However, DTU's algorithm had to be adjusted to be able to use input electricity data and had to be fine-tuned for different heating system types.

Chapter 4
WEATHER DATA

To more accurately predict a building's energy performance, the NEEM core solution aims to utilise weather data points from the closest meteorological stations. The NEEM core solution and DTU's prediction algorithm require outdoor air temperature, wind speed, and global radiation.

Several weather services throughout the Nordic region offer both historic weather data and forecast data. Since weather data is not GDPR sensitive, meteorological data is often publicly and easily accessible. This data can be found either via a file explorer and click-and-download solution, or via an open API service. The national weather solution was used for the Danish and Swedish NEEM pilots. In the Norwegian pilot, the weather data was based on the Copernicus data service. Copernicus is a European earth observation program that provides ground- and satellite-based weather measurements and forecasts. The Copernicus data service comprises every possible weather measurement in fine-grained resolution. However, retrieving data from Copernicus had its own limitation, namely that the extraction of data from the data formats commonly used in meteorology proved to be somewhat challenging. This challenge is described in more detail under the Norwegian pilot.

It is important to note that weather observations from different data sources may cause inconsistencies in the energy performance model predictions due to different parameter observation methodologies, if not considered appropriately. This concern can also be addressed by using the Copernicus data service. Furthermore, the Copernicus API solution would eliminate the need to create separate data ingestion pipelines for each Nordic country and its national meteorological institutes.

4.1 DENMARK

4.1.1 Data availability

In Denmark, weather parameters were retrieved from the Danish Meteorological Institute (DMI) data portal. The Danish Meteorological Institute provides the community with meteorological knowledge and data within the Commonwealth of Denmark, the Faroe Islands, Greenland, and the surrounding waters and airspace. The Danish Meteorological Institute's tasks cover weather, climate, and sea, ranging from issuing weather forecasts and warnings of dangerous weather to producing ice information for Greenland and developing future climate scenarios for use in climate adaptation in Denmark. For the NEEM core solution purposes, the weather data was retrieved using the DMI Open Data API.

4.1.2 Data access, processing, and challenges

After registering as a user on the DMI portal, it was possible to freely access DMI's data. Out of the available services, we used the meteorological observation (metObs) API that contains raw weather observation data, e.g. wind, temperature, and global radiation data, from DMI-owned stations.

4.1.3 Data quality

No significant data gaps were encountered when working with the weather parameter data from DMI. However, some of the closest DMI weather stations did not have global radiation parameters. This situation was solved by taking the data from the next closest weather station that had data related to global radiation.

Since the buildings in the Danish demonstration site contained coordinates (their longitude and latitude) that were obtained from static building data, it was possible to calculate the Haversine distance to select the closest DMI weather station.

4.2 NORWAY

4.2.1 Data availability

Weather parameters were retrieved using the Copernicus climate program and data was downloaded from the Copernicus Climate Change Service (C3S), the Climate Data Store (CDS), and, more precisely, from the ERA5 dataset.

4.2.2 Data access, processing, and challenges

To access weather data from the Copernicus CDS, it was necessary to create a user profile and retrieve an API access token to download the data.
The additional difficulty came from the fact that the available download file formats in CDS were in GRIB and NETCDF format. It was necessary to extract all the relevant data parts from these files and convert them into CSV files as necessary for the NEEM core solutions. The Center Denmark (CDK) data scientists had not encountered these file formats before and this created some additional challenges, which, in the end, were successfully solved.

4.2.3 Data quality

As the buildings did not have precise coordinates, the selection of the weather parameters was based on the postal code of each building. These postal codes were mapped onto the municipality code and weather parameters were selected based on the coordinates of the municipality area. No significant data gaps were encountered.

4.3 SWEDEN

4.3.1 Data availability

Weather data can be retrieved using the Swedish Meteorological and Hydrological Institute (SMHI) services. The Swedish Meteorological and Hydrological Institute is an expert agency under the Ministry of the Environment. The Swedish Meteorological and Hydrological Institute's observation stations collect large quantities of data, including temperature, precipitation, wind, air pressure, lightning, solar radiation, and ozone. The institute offers services to build applications using the SMHI Open Data API or by downloading files using explorer to select weather stations and weather parameters.

4.3.2 Data access, processing, and challenges

For a more scaled-up solution, a data ingestion setup with SMHI Open Data API should be used. However, for a small demonstration case, it was satisfactory to select the closest weather stations to the demonstration buildings and all relevant weather parameters manually.

4.3.3 Data quality

Some weather stations that contained outdoor air temperature and wind speed data did not have solar irradiation data; in these instances, it was necessary to select another weather station that contained the global radiation parameter. Afterwards, using multiple weather parameter files, the data were joined into a single table to match the DTU's algorithm input and the period of consumption data.

Chapter 5

DATA AVAILABILITY AND QUALITY ASSESSMENT SUMMARY

After describing all the necessary data to build the Minimum Viable Product (MVP) for the NEEM core solution by data type and country, it is important to look more closely at the data itself and summarise the findings. Having all the relevant data sources is essential for the NEEM core solution. Having high-quality data is even better, as this will allow the DTU algorithm of the NEEM core solution to estimate a building's energy performance with better accuracy. Even though the NEEM core solution is designed to handle some data gaps, the quality of the model's energy performance estimation may deteriorate if too many consumption periods are missing or the observation periods are too short in the heating season.

In this chapter, we take a closer look at the available data and its quality and further describe the processes used to select the data for each demonstration site.

5.1 CURRENT DATA AVAILABILITY AND GAPS

In this section, we provide a brief overview of the three Nordic countries and the currently available data sources across all data types. Data sources are mostly publicly owned/supported or come directly from energy suppliers. For some of the data, alternative sources are accessible for data comparison and quality assessment. If needed, the alternative data sources can act as a replacement dataset.

The weather data has the highest availability since it can be retrieved from both the national meteorological services in all aforementioned countries and also from the open EU climate data hub, Copernicus.

In Denmark, static building data can also be freely accessed as public data. On the other hand, in Norway and Sweden, a national electronic ID (eID) or digital signature solution is required to log in and query specific building data. Since the data preparation team had neither Swedish nor Norwegian eIDs, this solution was not possible. For these countries, the relevant static building data, i.e., area and the number of floors, was acquired from the utility companies as an alternative source.

Besides the utility companies, which are the main electricity consumption data providers, Denmark and Norway have public IT systems that can collect electricity production and consumption data and can grant third-party access-solution secure access to their data.

Table 1
Data availability in Denmark

Weather Data		
Temperature, wind speed, wind direction	Danish Meteorological Institute (dmi.dk)	Copernicus
Static Building Data		
Building information	Bygnings-og Boligregisteret (BBR)	
Energy label	Energistyrelsen's EMO data service (EMO)	
Energy Consumption Data		
Electricity	TREFOR El-net	
Heating	TREFOR Varme	Fredericia Fjernvarme

Source: Center Denmark

Table 2
Data availability in Norway

Weather Data		
Temperature, wind speed, wind direction	Norwegian Meteorological Institute (met.no)	Copernicus
Static Building Data		
Building information	Kartverket (kartverket.no)	
Energy label	Energimerking (Energimerking.no)	
Energy Consumption Data		
Electricity	Elhub (elhub.no)	ELVIA
Heating	Heating data is not being used for this NEEM demonstration site	

Source: Center Denmark

Table 3
Data availability in Sweden

Weather Data		
Temperature, wind speed, wind direction	Swedish Meteorological and Hydrological Institute (smhi.se)	Copernicus
Static Building Data		
Building information	Boverket (boverket.se)	
Energy label	Boverket (boverket.se)	
Energy Consumption Data		
Electricity	Stockholm Exergi/SEOM	
Heating	Stockholm Exergi	

Source: Center Denmark

5.2 DATA QUALITY AND GAPS

5.2.1 Quality and gap analysis approach

For the NEEM project, data quality and gaps are assessed using the following approach:

The *Source* column describes the data source. In the case of multiple data providers, the rows are split up and data gaps are described individually for each provider.

The *Quality Assessment* column describes the level of data quality. The assessment contains a subjective data evaluation ranging from HIGH, MEDIUM to LOW. This assessment is based on the data availability and observed data gaps.

The *Data Gaps Observed* column contains a percentage of the missing data of the available dataset, i.e., <5%, meaning that less than 5% of data is missing.

The *Measures Taken to Counter Data Gaps* column contains a brief description of the contagion plans and actions taken to counter large data gaps.

Building Energy Performance Certificate (EPC) labels are referred to as *Energy Labels*. Also, it is important to note that energy label data is optional and is mostly used to evaluate the NEEM core solution prediction accuracy. Additionally, it is important to mention that the energy-labelling scale is different across Nordic countries.

5.2.2 Quality and gap analysis: results by country

Table 4
Quality and gap analysis table, Denmark

Data Type	Source	Quality Assessment (High; Medium; Low)	Data Gaps (%)	Measures Taken to Counter Data Gaps
Static Building Data				
Floor size	BBR	High	0%	No action.
Location	BBR	High	0%	No action.
Heat source	BBR	High	0%	No action.
Energy label	EMO	Low	~50%	Data is optional. No action.
Energy Consumption Data				
Electricity consumption	TREFOR El-net	High	<5%	Data is optional for this demonstration site.
Heat consumption	TREFOR Varme	Medium	<10%	Improving data ingestion in collaboration with the utility company.
	Fredericia Fjernvarme	High	<5%	No action.
Weather Data				
Outdoor temperature	DMI	High	<1%	No action.
Wind speed	DMI	High	<1%	No action.
Wind direction	DMI	High	<1%	No action.
Global solar irradiation	DMI	High	<1%	Selecting the data from the next closest station if the parameter is missing.

Source: Center Denmark

Table 5
Quality and gap analysis table, Norway

Data Type	**Source**	**Quality Assessment (High; Medium; Low)**	**Data Gaps (%)**	**Measures Taken t Counter Data Gaps**
Static Building Data				
Floor size	Elvia	High	0%	No action.
Location	Elvia	High	0%	Using postal code instead of coordinate to retrieve loc weather data.
Heat source	-	-	-	-
Energy label	-	-	-	Data is optional. N action.
Energy Consumption Data				
Electricity consumption	Elvia	High	<1%	No action.
Heat consumption	-	-	-	Data is optional for th demonstration site. N action.
Weather Data				
Outdoor temperature	Copernicus	High	<1%	No action.
Wind speed	Copernicus	High	<1%	No action.
Wind direction	Copernicus	High	<1%	No action.
Global solar irradiation	Copernicus	High	<1%	No action.

Source: Center Denmark

Table 6
Quality and gap analysis table, Sweden

Data Type	Source	Quality Assessment (High; Medium; Low)	Data Gaps (%)	Measures Taken to Counter Data Gaps
Static Building Data				
Floor size	Swedbank/SEOM	High	0%	No action.
Location	Swedbank/SEOM	High	0%	Using postal codes instead of coordinates to retrieve local weather data.
Heat source	Swedbank/SEOM	High	0%	No action.
Energy label	-	-	-	Data is optional. No action.
Energy Consumption Data				
Electricity consumption	Swedbank/SEOM	High	<1%	No action.
Heat consumption	-	-	-	Data is optional for this demonstration site. No action.
Weather Data				
Outdoor temperature	SMHI	High	<1%	No action.
Wind speed	SMHI	High	<1%	No action.
Wind direction	SMHI	High	<1%	No action.
Global solar irradiation	SMHI	High	<1%	Selecting the data from the next closest station if the parameter is missing.

Source: Center Denmark

5.2.3 Structural data quality concerns

Besides the individual dataset data quality, such as missing data and availability, one must be aware of the structural issues that can be caused by not using the data appropriately despite the high individual data quality. We identified several significant concerns.

Firstly, one must be aware of the differences in the definition of weather parameters in different climate data providers. To illustrate this, some weather parameters available from the Danish Meteorological Institute (DMI) are different from the ones available at Copernicus Data Services (CDS). For example, the wind speed from the Danish Meteorological Institute is measured at 10m over terrain and, from Copernicus, this parameter is given as 10m u-component and 10m v-component parameters separately. These differences must be considered in the energy efficiency modelling tool setup to avoid inconsistencies.

Secondly, outdated building data information (from the Danish Building Registry Database) regarding the space heating system may lead to a faulty energy efficiency prediction. For example, not knowing that additional heating sources, such as heat pumps, have recently been installed in a building will lead to challenging predictions.

In cases where the building is heated by electricity, it is important to know what the heat source is (direct heating or a heat pump) and if the building owners have acquired an electric vehicle and are charging it at home. Such a situation can alter the model prediction accuracy if it is not considered appropriately.

5.3 DATA COLLECTION AND PROCESSING CHALLENGES

5.3.1 Data collection

For the NEEM core solution to be tested and run smoothly, it was necessary to gather data in one place. The Center Denmark data platform served this purpose. Many data collection challenges were encountered in the process.

The difficulties accessing data varied considerably depending on the data type. For example, the climate data in all Nordic countries was available as public data; it was fairly easy to access, and it could be integrated via data fetching APIs.

On the other hand, the consumption data appeared to pose the biggest challenge, since all the utility companies had their own data-sharing processes. Among the consumption data types, electricity data was the most readily available since many Nordic countries have centralised IT systems (Norway, Denmark) that gather smart meter data points in one place and thus facilitate data sharing. Furthermore, district heating data (in Denmark) provided an extra challenge since a centralised data hub does not exist for this type of data.

5.3.2 Data processing

After loading all the necessary data onto the CDK data platform, it had to be transformed into a format that the DTU energy prediction algorithm could consume.

The data processing tasks included dealing with a variety of data formats. First and foremost, time series data had to comply with the correct date format. Dealing with faulty and missing values necessitated extensive pre-processing of the data. Then, combining data from multiple data sources was another challenge, as the primary keys across the different data sources had to be additionally pre-processed. For example, the household address strings in the consumption and static building data tables had to be normalised to be eligible for a successful data merge.

Another challenge was locating the closest weather station for buildings that did not have coordinates. This was solved by relying on the zip and/or municipality codes.

It is important to mention another aspect related to data processing, namely the use of data processing tools and techniques. The dataset available in Denmark had a significantly higher number of households; thus, processing data efficiently involved the use of cluster solutions and big data processing tools like Apache Spark. On the other hand, the NEEM pilots in Norway and Sweden had less than 50 households, and data could be processed with more conventional data processing tools, like using Python Pandas Library on a single computer.

Chapter 6

A BRIEF PERSPECTIVE – WHY ARE THESE DATA IMPORTANT FOR FINANCIAL INSTITUTIONS?

To offer green mortgages to retrofit homes and make them energy efficient, financial institutions typically need access to a variety of data related to the property and the borrower. According to our research among financial institutions as partners and externally, we have come to the conclusion that banks find it difficult to fill in the data gap needed to provide new business models related to energy efficiency.

Here are the types of data that financial institutions deem necessary:

- Energy performance data: Financial institutions would need information about the energy efficiency of the property, which can be obtained through an energy audit or energy performance certificate. This data provides insights into how much energy a property is consuming and how it can be improved.

- Retrofit costs: Lenders would require estimates of the costs of retrofitting the property to make it more energy-efficient. These costs would include the cost of materials, labour, and other expenses.

- Property value: Financial institutions need to know the market value of the property to determine the mortgage loan amount.

- The borrower's financial information: Financial institutions would also need to obtain financial information from the borrower to assess their creditworthiness and determine the interest rate and other loan terms. This would include details about the borrower's income, employment, credit score, and other financial obligations.

- Government incentives and rebates: Financial institutions would also need to know if there are any government incentives or rebates available to support the retrofitting of the property. This could include federal, state, or local rebates, tax credits, or other incentives to support energy efficiency.

Overall, the data required for green mortgages is focused on assessing the value of the property, the cost and benefits of retrofitting it, and the borrower's financial capability to support the mortgage payments.

Lastly, financial institutions have expressed concerns regarding the data foundation needed to provide energy-efficient mortgages to retrofit homes. To offer such mortgages, lenders require reliable data to evaluate the energy performance of homes and the potential energy savings from retrofit measures.

However, accessing such data can be challenging, as it often requires extensive data collection and analysis from various sources, including utility companies, building owners, and energy auditors. In addition, the lack of standardised data and metrics for measuring energy efficiency can make it difficult to compare and evaluate the energy performance of homes. These challenges can hinder the development of energy-efficient mortgage products, as financial institutions require accurate and reliable data to assess the credit risk associated with lending for energy-efficient retrofits.

Chapter 7

CONCLUSION

Overall, the data quality is rather high if one looks separately at each of the data sources. The data sources with the highest data quality were the weather and climate data providers. If the local weather or national weather stations are not available, one can always turn to the European climate programme's Copernicus data service.

Depending on the energy consumption data provider, data quality and accessibility vary. In some situations, when the building is transferred to new owners or the building owners change their utility provider, data gaps can be observed. Some other data quality issues are related to the address string combinations. These arise when the consumption data and static building data are combined. In this case, address string normalisation is required.

At the same time, even though the static building data appeared to have no gaps, it is important to acknowledge that, on some occasions, it may be outdated. This is especially relevant when determining the heating source of the building. Also, no service currently exists that can inform the system if the users have an electric vehicle. Such information is especially critical for electrically heated buildings. This information must be obtained via questionnaires or inferred from the electricity data patterns.

This is especially common in Denmark, where the EPC labels are updated voluntarily or at the property sales point. Therefore, the EPC label data is available only for approximately half of the buildings and may also be outdated.
Additional barriers in Norway include the requirement of a national electronic ID to access building information and EPC data.

Overall, the biggest difficulty lies in the process of improving the accessibility of the energy consumption datasets. The biggest barrier in this case is the lack of digital access to anonymised consumer data. In Denmark, electricity consumption data is accessible through a centralised API, where each user can give access to their data to a third party. On the other hand, such a digital solution currently does not exist for the heating data. Therefore, this hinders the scalability of NEEM solutions in Denmark, as district heating is the main building heating source.

REFERENCES

Danish weather data service. Danish Meteorological Institute (DMI). https://www.dmi.dk/

API for Danish Energy Performance Certificates. Energistyrelsen's EMOData-service. https://emoweb.dk/emodata/api-docs/

Boligejer. Finding information and EPC about Danish properties. Danish Business Authority (Ehrvervsstyrelsen). https://boligejer.dk/

Meteorological Observation (metObs API) from the Danish Meteorological Institute (DMI). https://confluence.govcloud.dk/display/FDAPI/Meteorological+Observation

Swedish Meteorological and Hydrological Institute (SMHI). https://www.smhi.se/

SMHI Open Data API Documentation. http://opendata.smhi.se/apidocs/

Danish building information database – Bygnings-og Boligregisteret (BBR), accessible from Danmarks Adressers Web API (DAWA). https://dawadocs.dataforsyningen.dk/dok/bbr

Eloverblik. Electricity data portal provided by Energinet. https://eloverblik.dk/welcome

Copernicus. European Climate Data Program. https://www.copernicus.eu/en

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2018): ERA5 hourly data on single levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on 07-SEP-2022), 10.24381/cds.adbb2d47

Elhub. Norwegian electricity data hub. https://elhub.no/en/

Kartverket. The Norwegian Mapping Authority. https://www.kartverket.no/en

Boverket. Swedish National Board of Housing, Building and Planning. https://www.boverket.se/en/start/

Lantmäteriet. Swedish Cadastral and Land Registration Authority. https://www.lantmateriet.se/en/

Elvia. Norwegian utility company. Electricity Provider. https://www.elvia.no/

TREFOR El-net. Danish utility company. Electricity Provider. https://www.trefor.dk/elnet/

TREFOR Varme. Danish utility company. District Heating provider. https://www.trefor.dk/varme/

Fredericia Fjernvarme. Danish utility company. District Heating provider. https://www.fredericiafjernvarme.dk/

https://www.dingeo.dk/fjernvarme/trefor-kolding

https://www.dingeo.dk/fjernvarme/fredericia

Supply area of Fredericia Fjernvarme Utility Company. DinGeo – Simplified geographic data access. https://www.dingeo.dk/fjernvarme/fredericia

Supply area of TREFOR Varme Utility Company. DinGeo – Simplified geographic data access. https://www.dingeo.dk/fjernvarme/trefor-kolding