Plan Overview

A Data Management Plan created using DMPonline

Title: REDUCEDHEATCARB data collection winter 2023-2024

Creator: Henri ter Hofte

Principal Investigator: Henri ter Hofte

Data Manager: Henri ter Hofte, Nick van Ravenzwaaij, Iris Kiesling

Affiliation: hogeschool Windesheim

Template: DCC Template

Project abstract:

The main goal in the project REDUCEDHEATCARB is to design, implement and validate prototypes of data analysis algorithms that can be used as a basis to give advice to households with connected thermostats and/or connected boilers to reduce greenhouse gas emissions attributed to home heating, for their specific building, installation and comfort demand.

More in particular, we focus on an ordered list of sub-questions, answers to which might inform and/or motivate households to take interventions. These sub-questions and associated interventions are ordered from easy to more challenging interventions:

1. How much energy, CO_2 emissions and money can a specific household save after **changing the temperature setting and/or timing of their thermostat program**? (e.g., by lowering the non-weekend morning heating setpoint by 0.5 ˚C?)
2. To what extent can a specific household**limit the supply temperature and/or maximum heating power of the gas-fired boiler** in their home…
   - … without substantial losses in thermal comfort?
   - … and how much energy, CO_2 and money can they expect to save?
3. After switching from a gas-fired heating system to a**hybrid heating system** in their home by adding only a relatively small heat pump and keeping their existing gas-fired boiler …
   - … (to what extent) can that household expect the same thermal comfort?
   - … and how much energy, CO_2 and money can they expect to save?
4. After switching from a gas-fired heating system to an**all-electric heat pump**…
   - … (to what extent) can that household expect the same thermal comfort?
   - … and how much energy, CO_2 and money can they expect to save?

For each of these sub-questions, an additional question is also: what is the minimal set of data (in terms of duration, properties, precision, temporal resolution) required to still give a reliable estimate?

The purpose/outcome goals of the project extend beyond the development and validation of data analysis algorithms. The ultimate aim is to contribute to a significant reduction in greenhouse gas emissions attributed to home heating on a national scale.

ID: 131894

Start date: 01-06-2023

End date: 30-09-2024

Last modified: 03-11-2023

Grant number / URL: TFF2305

REDUCEDHEATCARB data collection winter 2023-2024

Data Collection

What data will you collect or create?

1. Via a recruitment survey
   (answers questions marked by * serve as inclusion/exclusion criteria:
   1. name + e-mail address + informed consent*
   2. that subject is a customer of Remeha (we recruit subjects via Remeha)
   3. home address*; this info used
      1. as the destination address to deliver measurement devices
      2. as an index to retrieve public building data from online sources such as PDOK, BAG, 3D BAG Viewer, EP-Online and public weather data such as KNMI Dataplatform
   4. properties of the home:
      1. type of dwelling (single-family dwelling / multi-family dwelling);
      2. whether the home has an internet connection and wireless internet (Wi-Fi)*;
      3. whether the home is heated by gas-fired heating boiler* and/or other means;
      4. whether gas-fired heating boiler employs floor heating to release heat;
   5. smart energy meter:
      1. whether the home has a smart energy meter*;
      2. make and model of the smart energy meter;
      3. full number of the smart energy meter.
   6. about the occupants:
      1. number of household members;
      2. whether any household member is 11 years old or younger*;
      3. whether any household member is 15 years old or younger;
      4. whether all household members have a smartphone;*
      5. number of household members for each smartphone operating system category (iOS / Android / other);*
      6. whether all household members are prepared to leave on Bluetooth during the measurement period;*
      7. informed consent from parents to include children between 12 and 16 years old in the occupancy count;
      8. whether the subject willing to receive a certified Energy Performance Certificate (EPC) inspector to perform an EPC inspection of their home;*
      9. whether subjects have an electric vehicle that is charged often at home;*
      10. whether people use gas or electricity to cook.
2. Via measurement devices we send to subjects and they install at their home:
   1. living room module:
      1. indoor climate monitoring data (indoor temperature, CO_2 concentration, relative humidity): several times per hour;
      2. occupancy data: the number of smartphones of regular occupants of the home that registered with the living room module and gave informed consent for their presence at home to be included in the count, as well as the actual number of such smartphones that respond to Bluetooth name requests (several times per hour).
   2. smart energy meter readings (gas, electricity) via the P1-port (several times per hour);
3. Via an EDSN-certified Overige Diensten Aanbieder (ODA): smart energy meter readings (gas, electricity) via the P3/P4 port
   1. 1. 15 minute interval electricity meter readings (up to 10 days in the past);
      2. 1 hour interval gas meter readings (up to 10 days in the past);
      3. daily meter readings (up to 40 days in the past);
      4. monthly meter readings (up to 13 months in the past).
4. Via data export by BDR Thermea of consenting subjects:
   1. boiler and thermostat data: which model boiler and thermostat does the subject have in their home;
   2. heating system monitoring data
      - high frequency data (~ every second)
        - indoor air temperature;
        - outdoor air temperature;
        - indoor heating temperature setpoint;
        - natural gas consumption intensity by the boiler;
        - whether the boiler is currently using gas for heating or domestic hot water;
        - setpoints for maximum hydronic supply water temperature and heating power;
        - hydronic supply and return water temperature;
        - hydronic flow rate;
        - setpoint for domestic hot water temperature;
        - domestic hot water temperature;
        - domestic hot water flow rate;
        - total amount of gas used by the boiler for heating;
        - total amount of gas used by the boiler for domestic hot water;
      - state data (state at beginning of measurement period and changes during measurement period)
        - thermostat program;

        - setpoint for maximum supply temperature;
        - setpoint for maximum heating power.
5. Via a certified EPC inspector: all data that is available in an EPC inspection, and used to calculate an energy label, including:
    1. type of home, type of roof, type of building mass;
    2. usable floor area;
    3. floor area(s): area and Rc value;
    4. roof area(s) area and Rc value;
    5. (exterior) walls: area and Rc value;
    6. windows: area and U value;
    7. doors: area and U-value;
    8. heating system(s): type, fuel, heat distribution system type, supply temperature category;
    9. cooling system: no / compression / ground cooling;
    10. tap water (type, CW-class);
    11. solar boiler (yes/no, surface area, orientation);
    12. ventilation (type, installation year);
    13. photovoltaic solar energy system (yes/no, surface area, type, orientation).

The volume of data will be several GB for tens of homes measured for several months. We dimensioned the relational database at our Twomes backend server such that it can handle this with ease.

How will the data be collected or created?

We collect data via the following means:

- **data under I above:**
  will be collected via online mobile survey hosted on https://windesheim.eu.qualtrics.com/
- **data under II above:**
  will be measured via automated measurement devices in the home, enabled by the NeedForHeat app (published by Windesheim in the Google Play Store and Apple App store of Windesheim), uploaded via a secure connection over home Wi-Fi and internet to servers at energietransitiewindesheim.nl, hosted by Strato AG
- **data under III above:**
  will be measured via an independent service provider ('ODA') certified byEnergie Data Services Nederland (EDSN) that subjects authorize to access their smart meter measurements) to which respondents give us access.
- **data under IV above:**
  will be measured by BDR Thermea, via their online connection to the boiler and thermostat of consenting customers. BDR Thermea will upload this data pseudonymized via a secure link to an online OneDrive folder of Windesheim with restricted access, which allows us to match the data with the proper subject.
- **data under V above:**
  will be observed by an certified EPC inspector.

Data quality is assured as follows; for:

- **data under I above:**
    - we use validated questions from theWoON survey wherever possible;
    - for brand and model of the installation devices, we allow users to select the most common brands and models.
- **data under II above:**
    - clocks on measurement devices are synchronized several times per day over the internet via NTP;
    - we always make a distinction between the value 0 (zero) and no data;
    - measurements are timestamped using the device clock (the smart meter device clock in case of smart meter data obtained via the P1 port;
    - just before measured data is uploaded, the upload is timestamped using the device clock and just after arrival at the server, the upload is timestamped using the server clock, making timestamps during analysis more robust against device clock inaccuracies (like time skew);
    - data is buffered in persistent memory of the measurement device whenever data fails to upload, and will be uploaded from persistent memory once connectivity is restored, thus making the data collection more robust to connectivity failures between measurement device and server;
    - each device records 'heartbeat' measurements (every 10 minutes) and uploads them every hour; heartbeats of measurement devices are monitored by back-end researchers and action is taken via the front-end helpdesk to signal respondents, when no heartbeats are received for days, thus making the data collection more robust to device failures and long term connectivity issues;
    - we currently do not plan to introduce validity checks before uploading measurement data, other than standard security measures against buffer overflow attacks;
- **data under III above:**
    - except for the 13 month monthly interval back-service, these measurements are redundant compared to data item II and can serve as repeated sample.
- **data under IV above:**
    - data is used from regular Remeha production systems; we receive production quality data.
- **data under V above,**
    - Only certified EPC inspectors will do the inspections.

Documentation and Metadata

What documentation and metadata will accompany the data?

After merging data from various streams and pre-processing timestamps (e.g. by converting Unix timestamps or UTC timestamps to fully qualified timestamps with local time in the Europe/Amsterdam plus a time zone offset), all raw measurements will consist of:

- id: a unique code of the home;
- source: the device type name of the measurement device, or 'KNMI' in case of weather data, 'ODA' in case of smart meter data collected via the P3/P4 interface, 'Remeha_<device>', for data originating from Remeha, where <device> will be replaced with an indication of the type of device that the data originated from.
- timestamp: a fully qualified timestamp (i.e. local time plus time zone offset)\
- property: property name of the measurement, which includes the unit of the value after a double underscore, using the 'physiquant__unit' variable naming convention
- value: value of the measurement

In a metadata table, we will describe, for each property:

- property name
- pandas datatype
- unit
- measurement interval [h:mm:ss]
- description
- source
- sensor
- format (using printf format specifiers from the C language)

Ethics and Legal Compliance

How will you manage any ethical issues?

1. Data collection plans are reviewed by the Windesheim ethics board
2. We inform subjects about the privacy policy repeatedly, which is continuously available via the measurement campaign website.
   This has a short summary as well as more complete coverage and we give information how to contact us (the front-end helpdesk, principal researcher or privacy officer of Windesheim) if they have questions
3. Informed consent will be asked from subjects:
   - by informing subjects about the privacy policy before they click on the online survey link;
   - by informing subjects again about the privacy policy at the beginning of the survey and ask explicit consent for data collection which starts with an online survey, which is recorded as one of the first answers to that survey, together with the name;
   - by informing them of the privacy policy again when they are invited to take part in the automated measurements;
   - for each of the measurement devices that subjects install with help of the NeedForHeat app, by shortly summarizing the data this particular measurement device collects, on the screen of the NeedForHeat app, after they scanned the QR-code of the measurement device with the app;
   - via a link displayed as a QR-code on the e-ink screen of the measurement devices, after the device has been configured with help of the app to start measurements.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

For hardware, software, data and reports created by Windesheim, we apply the following open licenses:

- **CERN-OHL-P v2** open hardware license to all hardware (designs) for measurement devices;
- **Apache License 2.0** open source license to all:
  - firmware for Windesheim's measurement devices;
  - source code for Windesheim's NeedForHeat app;
  - software for the Twomes API and backend

- **CC BY 4.0** copyright license to any open data and report that we publish.

These licenses all have a permissive nature (without strong or even weak reciprocal licensing obligations), which allows for re-use and adaptation in practice, even in a commercial context.
In cases where Windesheim is not at liberty to choose a license as it sees fit, we will seek approval by the project steering board to publish hardware, software, data and reports under the least restrictive licenses possible, compatible with commercial exploitation.
One example concerns situations where we cannot avoid re-using available components, source code and/or libraries that have an incompatible license, such as the GNU GPLv3 source code license. In that case, we seek to publish our work under the least restrictive license possible.
Another example may be use of proprietary analysis algorithms by BDR Thermea and/or Remeha, e.g., to compare the open algorithms developed in the project with the proprietary algorithms. We will seek to publish as much as possible, e.g., the comparison results, but may be restricted to only publishing the results of the open algorithms developed in the project.
All our open source software, open hardware designs and open data in the project will be published on GitHub under Research group Energy Transition at Windesheim (https://github.com/energietransitie)

Storage and Backup

How will the data be stored and backed up during the research?

We keep the following data separate, linked via a pseudonym (a unique, random number per subject)

1. directly identifiable data (such as name, street address, e-mail address)
2. indirectly identifiable data (all other data, which may include a pseudonym)

The online survey responses contain a combination of both directly and indirectly identifiable data (hosted by Qualtrics LLC, located in Provo, UT, USA).
The online survey responses will be exported and directly identifiable data will be replaced by a pseudonym. This pseudonymized data may also be stored in data stores provided by Office 365 services and OneDrive for Business services under the terms agreed by Windesheim with Microsoft, headquartered in Redmond, WA, USA.
Based on the street addresses, specific building data will be downloaded from sources such as PDOK, 3D BAG Viewer and EP-Online and stored as indirectly identifiable information along with the pseudonym.
During the measurement phase, only indirectly identifiable measurement data will be uploaded to energietransitiewindesheim.nl (hosted by Strato AG, located in Berlin, Germany). During phase 2, we will perform regular backups of measured data (several times per week) on SURFdrive (hosted by SURF B.V., located in Utrecht, the Netherlands).
Pseudonymized data collected by Remeha from consenting subjects will be uploaded via a secure connection to a folder on Microsoft OneDrive for Business. Access to this folder will be restricted.
EPC data collected by EPC inspectors will be uploaded via a secure connection to a folder on Microsoft OneDrive for Business. Access to this folder will be severely restricted.
Analysis will take place on JupyterLab Notebooks hosted on the server energietransitiewindesheim.nl (hosted by Strato AG, , located in Berlin, Germany), and/or local JupyterLab Notebooks hosted on researcher laptops.
Windesheim has data processing agreements with Qualtrics LLC, Strato AG, SURF B.V., and Microsoft.
Windesheim and BDR Thermea will sign a data exchange agreement that governs the exchange of data needed to transfer pseudonymized boiler and thermostat monitoring data from consenting Remeha customers from BDR Thermea to Windesheim.

How will you manage access and security?

Directly identifiable information will be stored in a separate 'key file', which links each pseudonym with directly identifiable information such as name, street address and e-mail address.
Access to directly identifiable information is severely restricted, i.e. to:

- selected staff at Windesheim University of Applied Sciences and Remeha (who are bound to confidentiality based on their employment contract). At Windesheim, this file will be encrypted and stored in a secure location. At Windesheim, the encryption key is only available to the principal researcher, one researcher responsible for data management and one person from support staff of the research group Energy transition. At BDR Thermea and Remeha, the key file and the enctryption key are only available to the principal researcher, a person responsible for recruiting subjects and a data scientist responsible for preparing the pseudonimized data uploads.
- helpdesk at Remeha knows which of their customers are subject in this research project, which allows them to forward subjects with questions regarding the research to the research helpdesk at Windesheim.
- staff at the selected organisation performing the EPC inspections.

Even though Windesheim is responsible for data processing, we strive to minimize directly identifiable available to Windesheim personnel as much as possible; only the selected key research staff at Windesheim indicated above will be involved in dealing with

activities that require access to directly identifiable information (e.g. processing informed consent survey data, sending measurement devices to research subjects, transferring contact data to the firm performing the EPC inspections, access to the uploaded EPC inspection data).
Data analysts only get access to pseudonymized data, not to directly identifiable information.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

An anonymized version of measured data will be prepared for publication as open data. However, it's important to note that complete anonymity might not be achievable due to the nature of the data and potential re-identification risks by parties holding complementary datasets.
The process includes:

- Ensuring the removal of all key files that directly link identifiable information with pseudonyms.
- Conducting a thorough assessment to verify that the remaining data cannot reasonably be linked to natural persons or individual residences.
- Applying additional anonymization techniques as necessary.
- Documenting metadata similar to the approach detailed in this example.

The anonymized data will be published as open data on platforms such as https://github.com/energietransitie and/or DANS KNAW under the CC BY 4.0 copyright license.To mitigate re-identification risks and address privacy concerns:

- Prior to the publication of anonymized data, Windesheim will take measures to ensure that the data cannot be directly traced back to specific individuals or residences. However, it's acknowledged that entities like Remeha and/or Enelogic might possess data that, when compared, could potentially lead to re-identification of specific individuals or residences.
- After the conclusion of the research project and before public release, all personal data will be purged. Depending on the assessment, this might involve severing links between pseudonyms and any directly identifying information.

Furthermore, a waiting period will precede the release of data as open data. During this period, participants have the opportunity to request Remeha and/or Enelogic to delete any 'fingerprint' data they've collected. This grace period ensures participants can confirm the removal of their data before the raw research data becomes publicly accessible. We aim to provide an adequate waiting period that offers participants the necessary assurances while considering the purpose of data publication and any processing procedures.

What is the long-term preservation plan for the dataset?

See the answer to the "Selection and Preservation" question.

Data Sharing

How will you share the data?

See the answer to the "Selection and Preservation" question.

Are any restrictions on data sharing required?

The default IPR clauses in TechForFuture project contract need to be honored.
For more details, see the answer to the question "How will you manage copyright and Intellectual Property Rights (IPR) issues?" in the section "Ethics and legal compliance"

Responsibilities and Resources

Who will be responsible for data management?

As principal researcher and project leader, Henri ter Hofte is responsible for implementing the DMP.
He will delegate various data management tasks to Nick van Ravenzwaaij and Iris Kiesling.
Carlos Mora Moreno from BDR Thermea will be responsible for data management tasks for this research project at BDR Thermea and Remeha.

What resources will you require to deliver your plan?

The resources are planned as part of the REDUCEDHEATCARB project, which is financed by TechForFuture, Windesheim and BDR Thermea.