PROGRAMME: H2020-LC-SC3-2020-EC-ES-SCC
START OF PROJECT: 01.10.2020
DURATION: 48 MONTHS

DELIVERABLE 4.1: INDIVIDUAL ENERGY DEMAND FORECAST MODELS

Authors: KUL, BCM
Due date of deliverable: 30.06.2023
Actual submission date: 30.06.2023

Deliverable Name	Individual energy demand forecast models
Deliverable Number	D 4.1
Work Package	WP4
Associated Task	T 4.1
Covered Period	M01-M32
Due Date	30.06.2023
Completion Date	30.06.2023
Submission Date	30.06.2023
Deliverable Lead Partner	KUL
Deliverable Authors	KUL, BCM
Version	3.0

DISSEMINATION LEVEL

PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

Change Control

DOCUMENT HISTORY

Version	Date	Change History	Author(s)	Organisation
1.0	01.06.2022	Table of content drafted	KUL, BCM	KUL, BCM
2.0	16.06.2023	Document drafted	KUL, BCM	KUL, BCM
3.0	30.06.2023	Last version	KUL, BCM	KUL, BCM

DISTRIBUTION LIST

Date	Issue	Group
01.06.2022	Revision table of content	WP2 leader and WP4 contributor
16.06.2023	Revision document	WP2 (AYESA & IDENER) and WP4 contributor
30.06.2023	Submission and distribution to partners	PO + All partners

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

1 TABLE OF CONTENT

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

2 ABBREVIATIONS

Short Name	Meaning
B2B	Business to Business
B2C	Business to Customer
DSO	Distribution System Operator
W	Watt
PV	Photovoltaic
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
API	Application Programming Interface

3 EXECUTIVE SUMMARY

The goal of this deliverable is to collect the necessary information to develop energy demand forecast models. Energy demand forecast models are used in order to forecast how much power is consumed by the end-users during different moments of the day. This allows to optimise energy flexibility at an individual scale which is the main goal of work package 5 of the Islander project). To be able to forecast end-users' energy consumption adequately, we needed information about their past consumption (at least 1 year of historical data), current consumption, and behavioural habits. That information is linked together such that it gives insights on how different kind of end-users are consuming energy on a daily basis. This provides information on how to optimise the energy stored in the battery and the energy produced by the PV panels.

The behavioural analyses teach that in general end-users consume most energy during the middle of the week, and during the morning and evening. This is caused by different activities such as working, watching TV, and doing laundry. By combining end-users' demographics and energy consumption reporting, we try to create profiles of end-users to get insights into how various kind of end-users are consuming energy differently. However, as the analyses are based on missing data (data collection is not finished yet), the end-users' profiles are still preliminary.

Developing accurate forecast models required to firstly collect historical consumption data, clean it and analyse it. No consumption data from the end-users of Borkum were available, therefore, we had to find other sources of data, which was challenging. Secondly, multiple model approaches were tested, and results were analysed. The model providing higher accuracy was selected.

In the course of 2023, smart meters will be installed on Borkum. This will allow to collect data on end-users' real-time energy consumption. This will be matched with end-users' reporting of energy consumption and linked to the behavioural and forecast data. The result will be a precise forecast model predicting the energy consumption of the end-users. This will be possible only after one year of data collection so that the model can be trained on historical data.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

4 INTRODUCTION

An important part of the Islander project is to be able to predict the energy needs and production of the Borkum island. Energy needs and production are determined by different factors, such as the weather, energy prices, but also the end-users' energy demand and supply. This deliverable focuses on the forecasting of individual energy demand of Borkum's end-users.

The goal of the current deliverable is to develop an effective forecasting model. To do so, we collected different kind of data. To get insights into the island's energy consumption demand we collected power consumed data from Borkum businesses and French households. As the household data from Borkum were not available, we decided to base our first version of the model on French household data. The model will be adapted once the Borkum data are available. Besides of energy consumption data, we needed data on the kind of end-users present on the island. This would help us to match end-users' profile with their energy demand in order to predict energy consumption for different kind of end-users based on their profile. To do so, we collected data on behavioural habits and demographical data.

By matching those different data together, we aimed at developing profiles of end-users and their typical energy consumption pattern. KUL was responsible for the collection of behavioural and demographical data, which will be used to characterise different types of end-users. BCM was responsible for developing a model that forecast consumption data.

In the next section, the context and objectives of D4.1 are shortly defined. Thereafter, the different data collected are described. In the following section, the methods to clean and analyse the data are presented. In the last section, the different results are presented and interpreted. The report ends with a short conclusion.

It is worth noticing that the current deliverable is still preliminary, as we are waiting for the installation of the smart meters on the island to be able to test forecasting model on end-users' data. The results reported in this report are thus subject to changes (it is required to have at least one year of historical data to calibrate the models and measure the accuracy).

5 CONTEXT AND OBJECTIVES

One of the goals of the Islander project is to develop and integrate a smart IT platform on Borkum. Different features will be implemented into the smart IT platform, and one of those features is forecasting services. Optimizing energy flows at the end-user level will be performed by the smart IT platform. It will take the predicted energy consumption, the energy generation, the battery status, and the energy price as inputs. Those forecasting services are developed within WP4. In the current deliverable, we will deep-dive into the development of the individual energy demand forecasting service.

The objective of this module is to develop a model to forecast the energy demand of the 30 households and 3 businesses that are recruited in the Islander project, based on their attributes and their energy consumption behavior. The goal will be to develop end-user segmentation based on end-users' energy consumption habits, behavioral data, and demographics.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

6 DATA (METHODS & ANALYSIS)

This section is divided in 2 parts, one section dedicated to understanding the energy consumption habits of Borkum's end-users (paragraph 6.1). A social behaviour study was conducted using questionnaires.

The second section is dedicated to short-term forecasting of the consumption of households and businesses (paragraph 6.2). Multiple forecasting techniques were tested to find the best forecasting accuracy.

6.1 Island's end-users' energy habits

6.1.1 General information of the questionnaires

In order to get clear and detailed insights into the end-users participating in the Islander project, we collected different kind of data using different methods and during multiple data collection periods. In the Islander project, we aim to get an overview of two types of end-users present on the island, households, and commercial buildings.

The goal of collecting data on the island's end-users is to understand what kind of end-users are living on the island, how and how much they consume energy. With that information, we will be able to develop energy consumption profiles for each end-user on the island. Those energy consumption profiles will allow us to predict the future energy consumption of an end-user based on their profiles.

Three online questionnaires were used in order to collect the necessary information about the end-users:

1. Commercial building profile questionnaire used to collect data about the different commercial buildings on the island, their sector of activities, their energy needs, and appliances use (based on Statistics Canada, 2021 [1]).
2. Household profile questionnaire used to collect data on the different households, their habitation, appliances, and habitudes (based on US Energy Information Administration, n.d. [2]).
3. Diary survey sent for one week (repeated once every season) to the different households to collect data on their daily energy consumption (based on US Census Bureau, 2022 [3]).

6.1.2 Questionnaires' data collection procedure

The data were collected through the mean of an online questionnaire, computed on the platform Qualtrics. The end-users (households and commercials) received an e-mail asking them to fill in the questionnaire. In this e-mail, the end-users received general information about the questionnaire, their personal ID, and the hyperlink redirecting them to the correct online questionnaire.

Three questionnaires (see above) were used for the data collection: (1) commercial buildings profile questionnaire, (2) household profile questionnaire, and (3) diary survey.

On 27th February 2023, an e-mail was sent to the three commercial end-users in order to fill-in the commercial buildings profile questionnaire. On the same day, the 30 households participating in the project were also contacted in order to fill in the household profile questionnaire.

A few weeks later, the first wave of the diary survey's data collection started. From Monday 3 April 2023 until Friday 7 April 2023, a daily reminder was sent to the household end-users in

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

order to complete the diary survey. This was repeated from Saturday 15 till Sunday 16 April 2023 in order to gain insights in end-users energy consumption during weekends. This first wave will provide us with insights of end-users' energy consumption during spring. In order to gain insights in the energy consumption during summer, fall, and winter, the data collection will be repeated in August, October, and January.

6.1.3 Questionnaire data analysis

The sample used for the data collection consists of 3 commercial end-users and 30 household end-users. We collected data from the commercial buildings profile questionnaire, the household profile questionnaire, and the diary survey. In the following section, we will provide an extended description of the sample from which we were able to collect data. This will be done by reporting on the descriptive analysis of the commercial buildings and household profile questionnaires. In a later section (8.2), we will try to provide in-depth insights of how different kind of households are consuming energy.

6.1.3.1 Commercial buildings profile questionnaire

The goal of the commercial buildings profile questionnaire was to gain insights in the different kind of commercial buildings on the island. Information on the buildings' activities, date of construction, number of appliances, number of employees etc were used in order to create profiles. The three commercial end-users filled-in the questionnaire. Those three buildings are quite old, as all three were built before 1960. Two of the commercial end-users had only one activity in their building, while another one had three activities in the building. Building 1 has "recreation centre", "industrial", and "other" as main activities, building 2 has "other" as main activity, and building 3 reported "hotel" as main activity. The surface of building 1 (1528 m2) and building 2 (1333 m2) is -larger than the surface of building 3 (650 m2). Both building 1 and 2 report a similar number of hours of operation in the building (40 and 39 hours respectively), however building 3 reports a much larger number of operating hours (70 hours). This could be explained by the fact that an hotel has more extended opening hours than other buildings. The three buildings differ in the number of employees active in the building. Nine employees are active in building 1, 22 employees in building 2, and only two employees in building 3.

The descriptive analysis of the commercial buildings allows to conclude that the three buildings differ to some extent (surface, number employees, and number of operating hours). The differences could be explained by the fact that the three buildings are active in different sectors. In a next section (8.2.1) we will explore the differences in energy consumption between the three buildings.

6.1.3.2 Household profile questionnaire

The data collected with the household profile questionnaire are detailed and provide insights on different aspects of the end-users' situation. We were able to collect general descriptive data about the end-users (age, gender…), data about their home, their general behaviour and opinions, and their investments in green solutions (e.g., solar panels). With the diary survey, we were able to grasp insights on how their energy consuming activities were divided across the week. In the current section, we will present general descriptive insights about the end-user's sample, and in a later section (8.2.2) we will investigate whether it is possible to create different profiles of end-users based on their characteristics and energy consumption pattern.

6.1.3.2.1 Demographical characteristics

In total, 26 end-users filled in the complete household profile questionnaire. The mean age

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

was 52 years (*sd* = 9.53), with a range of 33 years to 68 years old. From the 20 end-users who filled in the questionnaire, 18 were men and 8 women. 18 end-users were working full-time, 4 were working part-time, and 4 others were retired. From the 22 end-users employed, 8 reports having at least one person at home during the day for three days a week, 3 reports having someone during the day at home for four days a week, 2 reports having at least one person in the house during the day for five days a week, and 9 reports having someone at home six days a week. This suggest that even in employed end-users, there is often someone at home, which can result in energy consumption. 11 end-users reported a gross yearly income of 25.000-50.000€, 10 reported a gross yearly income of 20.000-25.000€, 3 reported a yearly gross income of 15.000-20.000€, and 2 reported a gross yearly income of 10.000-15.000€. End-users were also asked about the number of individuals living in their house, 3 end-users reported living with four or more individuals, 5 reported living with three individuals, 14 reported living with two individuals, and 4 reported living alone. Thirteen end-users are owners of their home, one of them is renting, and the twelve remaining are occupying their home without paying a rent.

6.1.3.2.2 Home's characteristics

When investigating the kind of home in which the end-users are living, we see that thirteen end-users have a single-family house detached from any other house while the remaining thirteen end-users have a single-family house attached to one or more other houses. The mean house surface 120.47 m2 (*sd* = 27.64), with a range of 80 to 180 m2. Most of the houses (N = 14) are south oriented, only one was north oriented, another one was north-east oriented, four were west oriented, another four were south-west oriented, and two were south-east oriented. When investigating the way of heating their houses, most of the end-users (N = 24) heat their houses with a central furnace, while the remaining two are using a steam or hot water system with radiators or pipes and a built-in room heater burning gas, oil, or kerosene respectively. None of the end-users reported using air-conditioning in their house.

Different appliances can be present in a home. In the questionnaire, end-users were asked to mention how many appliances they have for the following house's parts: kitchen, living room, cleaning, office, and varia. The mean total number of appliances across the whole house is 43.08 (*sd* = 17.92), with a range of 0 to 61 appliances around the house. However, it might be worth mentioning that 0 appliances is much unlikely and must be an error of the end-user. The appliances that were more often reported by the end-users were: small kitchen appliances (N = 91), TV (N = 77), smartphone (N = 73), vacuum (N = 58), fridge (N = 50, washing machine (N = 50), and dryer (N = 49). This suggest that in general, end-users have utility appliances (small appliances, fridge…), cleaning appliances, and smartphones in their home.

6.1.3.2.3 Behavioural insights

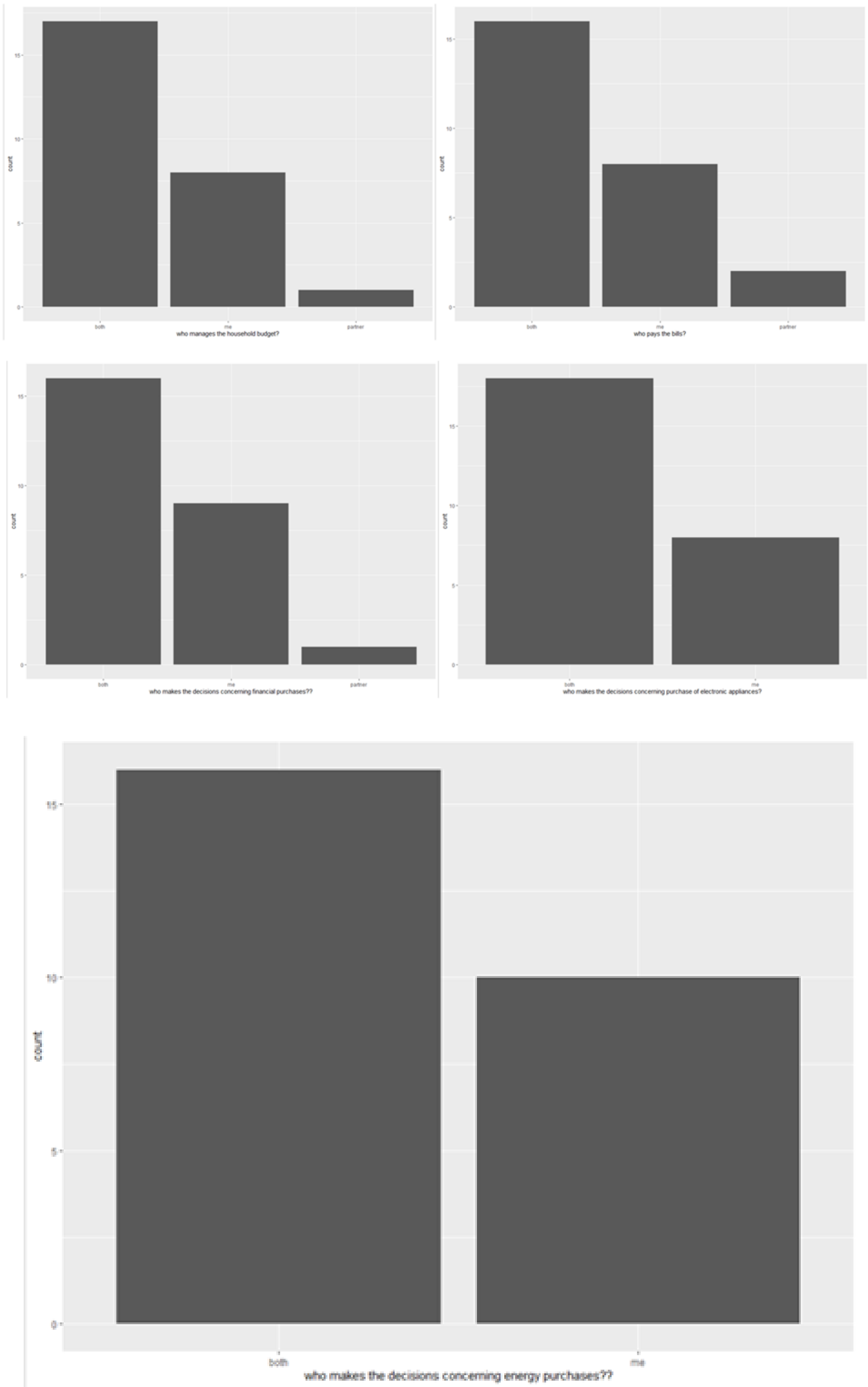
We were able to investigate different general behavioural pattern in the end-users. For this section, we will focus on the following insights: how end-users are using their thermostat for heating, how end-users and their household divide the management of economical and energy decisions, and how their behaviour changed due to the covid-19 crisis.

End-users were asked how they use their thermostat and how their temperature is regulated during winter and summer. Four end-users reported not having a thermostat, eight reported having a non-programmable thermostat, three end-users reported having a smart thermostat, and the majority (N = 11) reported having a programmable thermostat. When asking about how temperature is controlled during winter, one end-user reported not having control over the temperature, six end-users select one temperature and don't adapt, eleven end-users manually adapt temperature as needed, one end-user turn on and off manually when necessary, and seven end-users have the smart thermostat controlling the temperature manually. A slightly different behaviour is reported during summer. Only three end-users select one temperature and don't adapt, three adjust temperature manually, only five let the smart thermostat control the temperature, eight end-users turn on and off as necessary, and

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

four reports no control over the thermostat. The descriptive results suggest that in general, most end-users try to control temperature in winter and adapt to the needs, while it is less striking during summer.

End-users were asked about who managed the household budget, who paid the bill, who made the decisions regarding the purchase of electronic products, financial products, and energy products. As showed on Figure 1 below, in most of the households both the end-user and partner are making the decisions together.



This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

Figure 1 Decisions regarding different questions in households.

At the end of the questionnaire, end-users were asked whether they had the feeling that their household experienced changes in daily life due to the covid-19 situation. Twenty-two end-users said no, and four said yes. From the four end-users reporting a change in daily life due to covid-19, three provided an explanation and mentioned an increase in home working and time at home. End-users were also asked whether they had the feeling that their household experienced a change in energy consumption due to the covid-19 situation. Three end-users reported a change in energy consumption, the remaining twenty-three didn't experience a change in energy consumption. The three end-users reporting a change in energy consumption agreed on the fact that the covid-19 situation resulted in an increase in energy consumption.

6.1.3.2.4 Green investments

End-users were asked about their potential engagement in consuming energy in more pro-environmental way. When asked whether they were willing to adapt energy consumption to become more pro-environmental, the majority (N = 23) said yes and three didn't know. When asked whether they find that they are consuming energy in an environmentally friendly way, 15 end-users said yes and 11 said sometimes. End-users were also asked whether they already invested in actions to reduce energy consumption, the majority (N = 20) said yes, and the remaining 6 end-users said no. Based on the end-users' answers on different questions of the questionnaire, we were able to derive some of the actions they took to consume in a more environmentally friendly way and/or to reduce their consumption. Only one end-user has solar panels installed, however this end-user is not sure whether his/her household's consumption behaviour was adapted to the production of the solar panels. Four end-users invested in a greener, electrical vehicle. All end-users have double-pane (N = 16) or triple-pane (N = 10) glass instead of the more consuming single-pane glass. Only three end-users report having a poorly insulated house. The remaining end-users report a well (N = 12) and adequately (N = 11) insulated house.

6.1.3.2.5 Weekly energy consuming activities

With the diary survey, end-users were asked to report on the activities they conducted the last 24 hours, and this was repeated each day for one week. End-users were asked to report about household activities requiring appliances, resulting in energy consumption. To gain insights into end-users' energy consumption at home, we only selected reports for which participants were at home when doing the activities. It is worth mentioning that only two end-users completed the diary survey for seven days straight, therefore the results described in the current section are based on incomplete reports of end-users. The most reported activities were sleeping, personal grooming, watching TV, working on a job other than main job, washing and drying laundry. As can be seen on Figure 2, end-users report a larger number of activities during their days during the middle of the week (Tuesday-Friday).

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

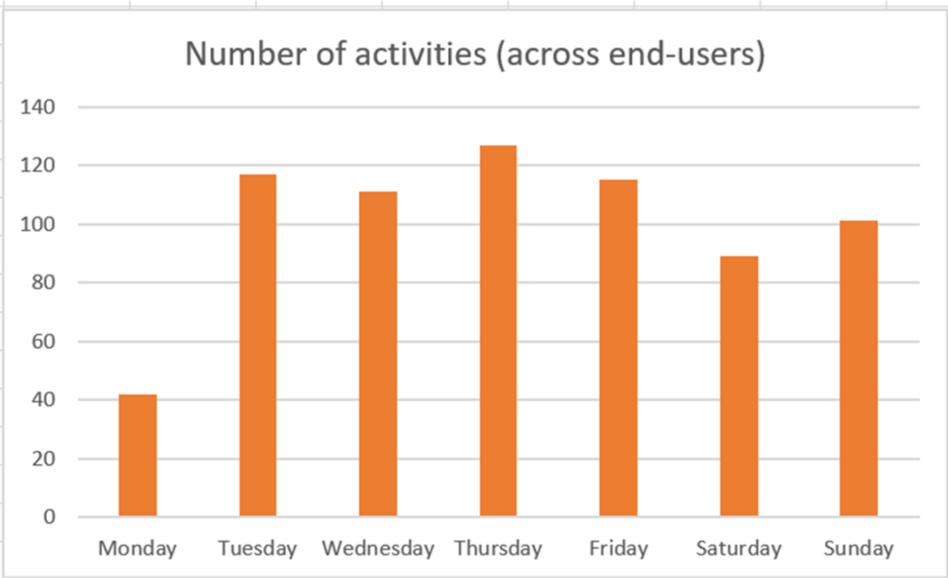


Figure 2 Number of activities across end-users for each day of the week.

When looking at the different moments of the day in which activities are reported (Figure 3), we see that most activities are spread across the day, with a lower amount of activities performed during lunch and night (during weekends).

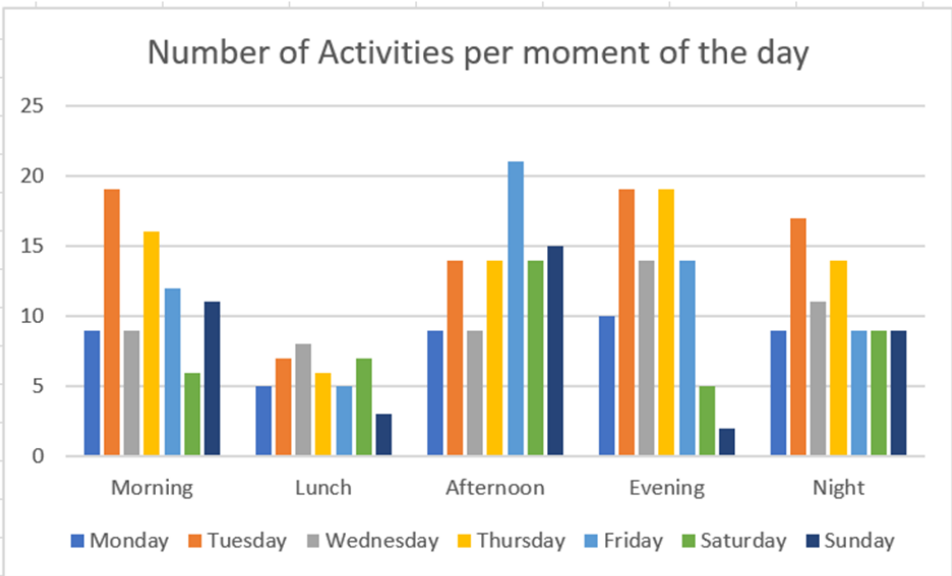


Figure 3 Number of activities per moment of the day.

6.2 Consumption data from households and businesses

One of the aims of the task is to develop a model that predicts the power consumed by the end-user of the Islander project. Two types of end-users are involved in the project: 30 household and 3 businesses. The end-user recruitment process is done, however; the PV panels and the smart meter that collect real-time observational data of consumption are not installed yet. Therefore, no real-time or historical data is available from the recruited end-users.

Data used to test the models was collected from 2 sources:

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

- Borkum utility collected consumption data at a granularity of 15 minutes of companies in Borkum (B2B) from 2018 to 2020.
- Anonymous B2C load curves from consumers in France at an hourly granularity for the year 2021.

The approach to build the forecasting model was the following:

- Finding, collecting, and choosing the relevant load curves.
- Cleaning and analysing the consumption pattern.
- Test multiple model approaches, analysing forecast accuracy on historical data.
- Finding the model providing the higher accuracy and testing several combinations of hyperparameters to improve the model. A hyperparameter is parameter that can influence the learning process which could improve model accuracy.

In the following paragraphs, the process will be detailed for the household data and business data. Each section will be divided in 3 parts:

- General information on the data.
- Data cleaning and load curve selection.
- Data analysis.

6.2.1 Commercial end-users

6.2.1.1 General information on the data

In the Islander project (WP2), three businesses have been selected to get PV panels, with the following activities:

- Sports and Youth Guest House: youth hostel with accommodation, kitchen, sports centre, and pitches.
- Office & Workshop: site with offices and craftsman's workshop.
- Living & Tourism: It's a bigger house with 6 holiday apartments.

Currently and until smart meters will be installed, there is no data of consumption available for these three businesses. Nevertheless, the consumption data of several Borkum businesses have been collected for the project, such as restaurants, hostels, shops, or other public installations.
Four years of consumption data from around 10 businesses were collected from Excel files covering the years 2017, 2018, 2019 and 2020. The metering point of the building and the power consumed in kW at a fourteen- or fifteen-minutes granularity is detailed.

6.2.1.2 Data cleaning

The data cleaning process was extensive because:

1. Lack of information on how data was measured.
2. Issues on the time zone (data UTC versus local time versus a mix of them).
3. Data consistency (issues on the order of magnitude of the consumption from a year to another)
4. Few information was available on the data (exact type of business, impact of covid on the business, etc.).

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

More details can be found in Annex 1 of the document and in Table 1.

Issue	Assumptions / actions
Lack of information on how data was measured.	It is assumed that the power consumed at a given time step is the mean power consumed in the fifteen previous minutes (ending convention). To keep the same granularity, each time step is also rounded to the matching quarter of an hour.
Issues on the time zone (data UTC versus local time versus a mix of them).	All the consumption load curves have been harmonized with local hours only (data from 2017-2020 was in local hour, whereas data from 2020 was in UTC)
Data consistency (scale issue).	Multiplication by 4 for all the data of 2020.

Table 1 Data cleaning actions and assumptions.

6.2.1.3 Load curve selection

A load curve is a plot of the power consumed over time (Figure 4 for example).

An analysis done on different load curves of businesses having the same activity highlighted similar trends and seasonality:

- For most of the businesses which provide accommodation services or food sale services, a consumption drop can be noticed in winter.
- For several of them, an exception occurred around Christmas until the first days of the year when a consumption peak can be observed.

Thus, data analysis methods and model calibration approaches will be performed on the youth hostel whose metering point is 5137, as a reference for the consumption of the prosumer "Sports and Youth Guest House ", which has the same business activity (Figure 4). For the other activities, no similar load curve was available.

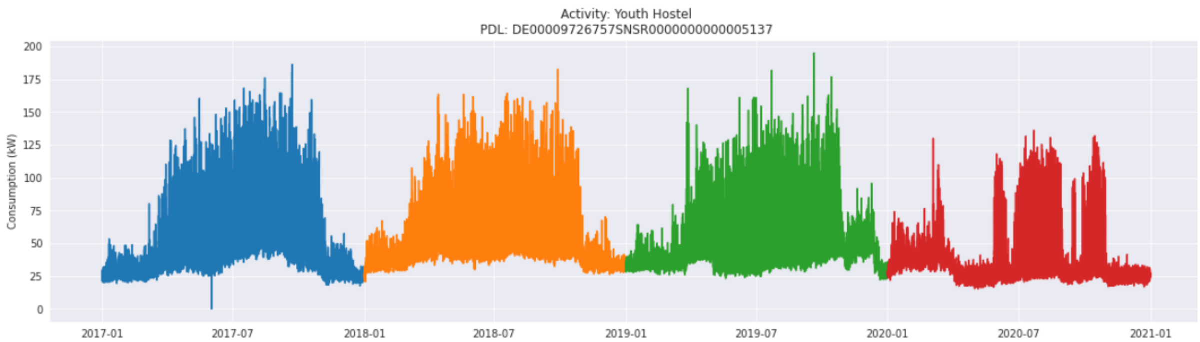


Figure 4 Load curve of the hostel with the metering point 5137.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

6.2.1.4 Data analysis

6.2.1.4.1 General analysis

During the three first years, seasonals behaviours and trends can be noticed (Figure 5):

- The consumption of the youth hostel is significantly lower in winter.
- The consumption increased until the summer.
- Lots of consumption peaks occurred, with a maximal threshold of 200 kW.
- Consumption never fell below 20 kW.

In 2020, the youth hostel consumption decreased significantly. A drop can be noticed between the end of March and the end of May 2020, at the end of June 2020, during two large periods in September 2020, and in November and December 2020. On the contrary, youth hostel consumption reached higher peaks than the other years in the beginning of March 2020.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

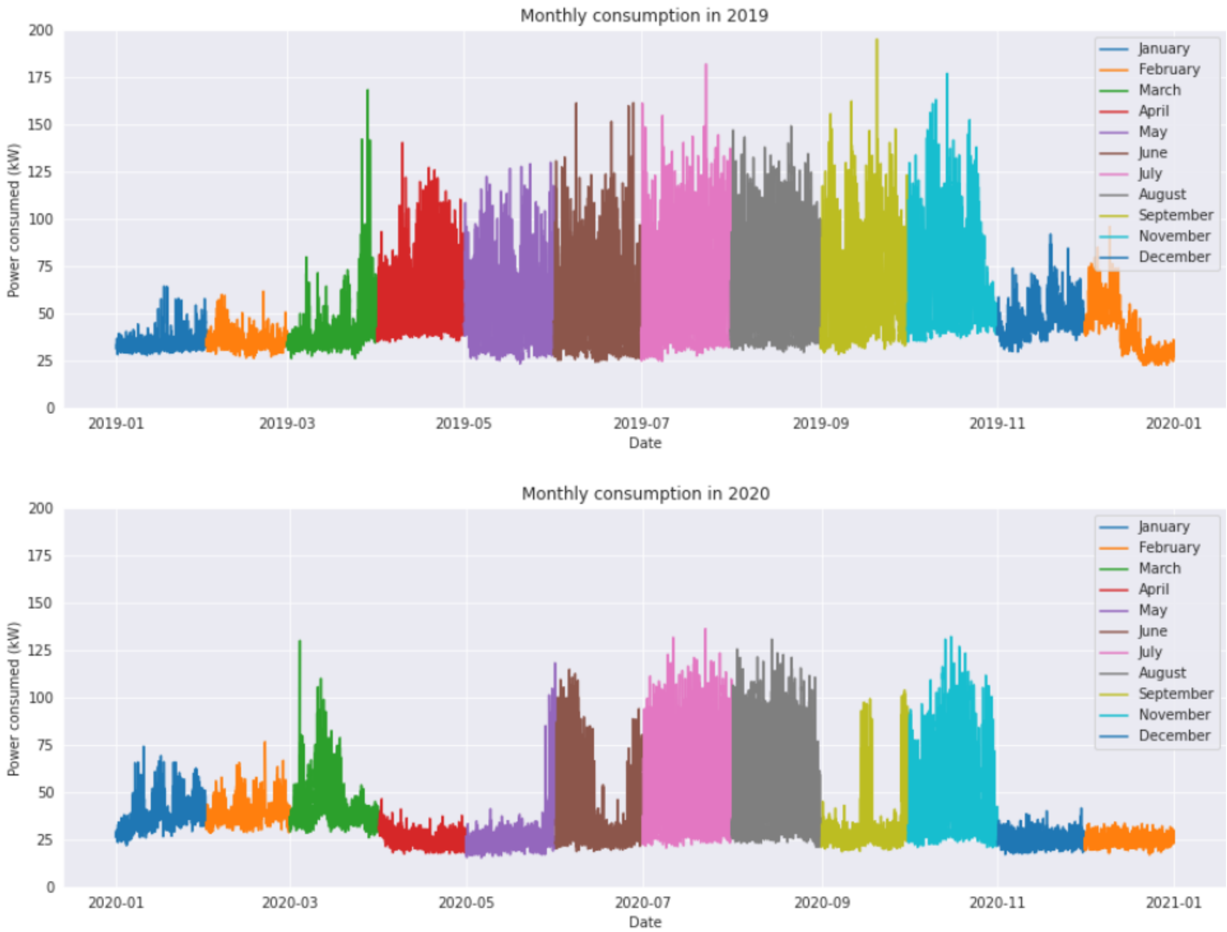


Figure 5 Monthly consumption of the hotel plotted by year.

The drops observed in 2020 can be explained as consequences of the lock downs, curfews and restrictions imposed to the island during the covid situation. This correlation between the covid pandemic and consumption is highlighted by the Figure 6, where an indicator of the stringency of the restrictions imposed on Borkum island has been plotted in black. Most of the time, an increase of the indicator matched to a drop of the consumption, as in March and in November.

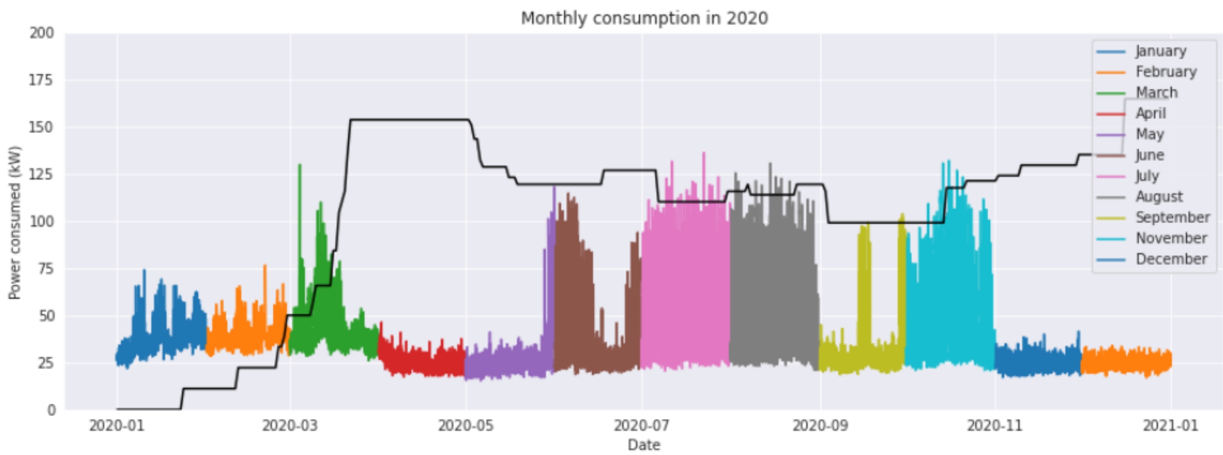


Figure 6 Monthly consumption of 2020 (colorful lines) correlated with covid stringency (black line).

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

6.2.1.4.2 Seasonality analysis

The paragraph below will focus on load curve seasonality at multiple granularities.

Monthly seasonality

The youth hostel consumption is each year higher in summer (Figure 7). This may be explained both by higher temperatures in summer that requires cooling and by the increased number of tourists booking a room in the hostel. In 2020, the consumption has been reduced due to the restrictions to face the covid.

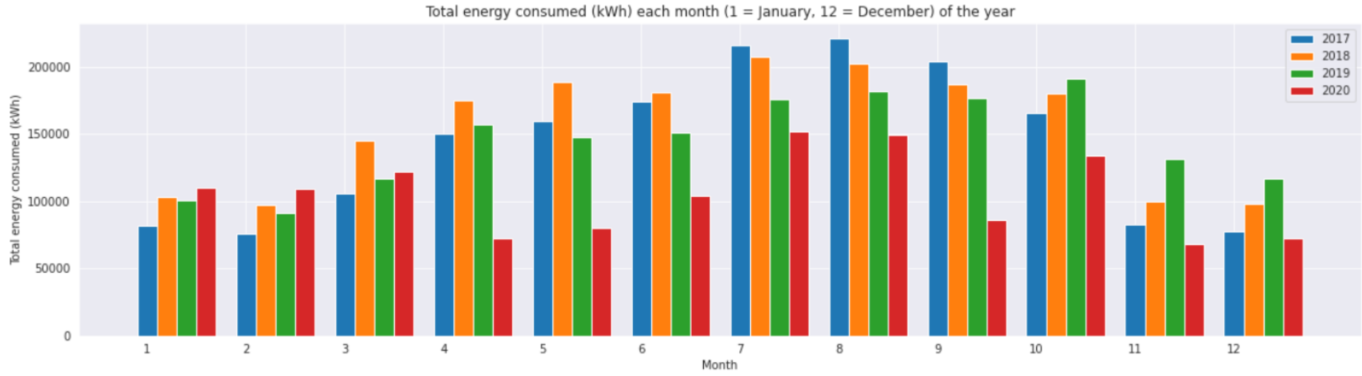


Figure 7 Monthly consumption of the years 2017 to 2020 expressed in kWh per month.

Daily seasonality

The daily energy consumed is lower during weekends, with highest values on Tuesday (number 1 in Figure 8) and Wednesday (number 2 in Figure 8).

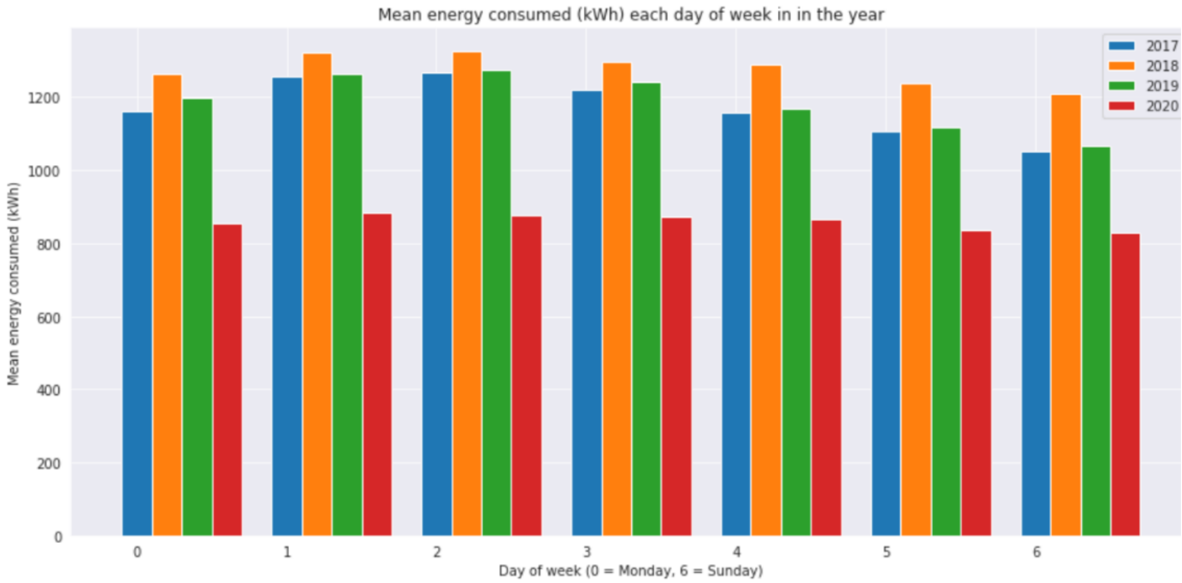


Figure 8 Mean daily consumption during the years 2017 to 2020 expressed in kWh per day.

Hourly seasonality

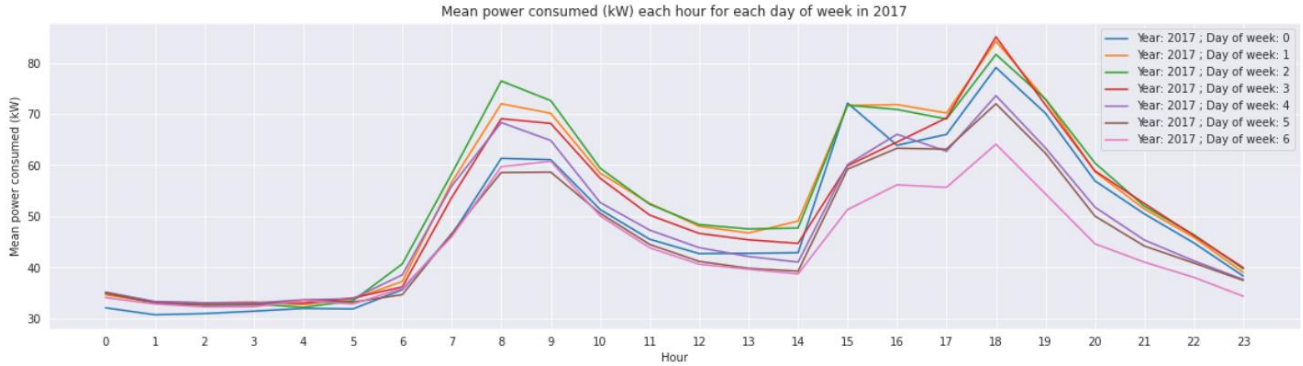
In Figure 9, the mean power consumed for each hour of the day has been computed for the whole year.

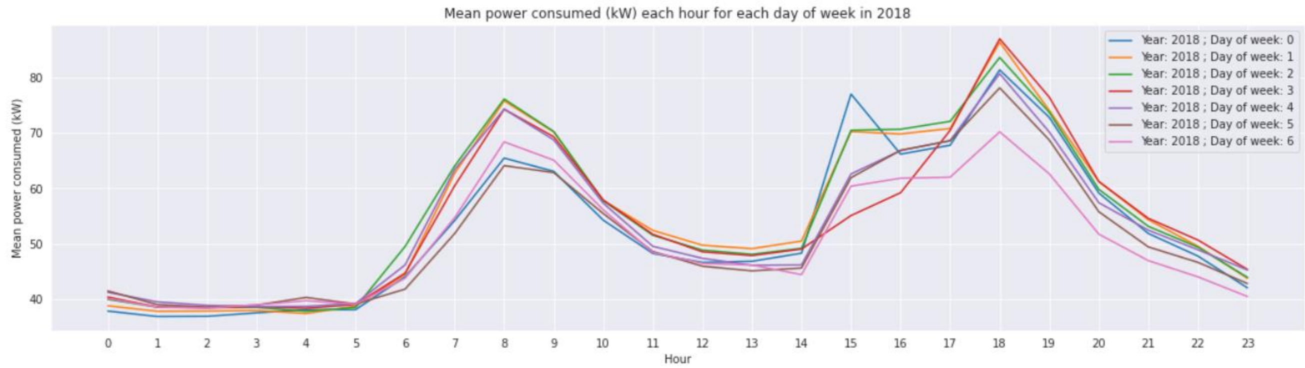
During 2017, 2018 and 2019, consumption behaviour was quite regular during the week. A consumption peak can be noticed in the evening from 6:00 p.m. to 7:00 p.m. This peak was

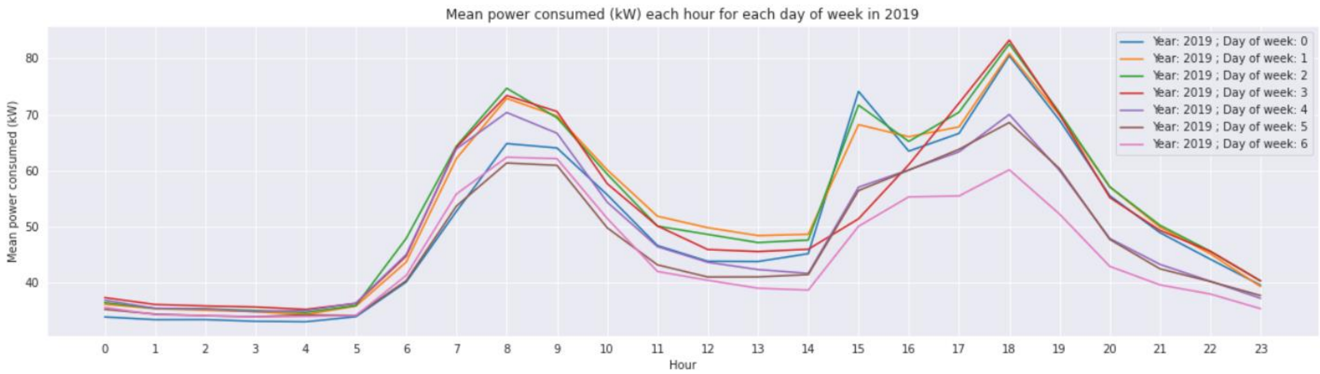
This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

higher on Tuesday, Wednesday, and Thursday, and much lower during the weekend. Before this peak, an increase of the consumption could be observed from 2:00 p.m. to 3:00 p.m. For a hostel, it is quite expected since more guests are presents in the evening than during the day, and this peak is matching with cooking hours. A second peak occurred from 8:00 a.m. to 10:00 a.m. This could also match with cooking activities for guests' breakfast or rooms cleaning, which consume energy.

In 2020, as expected, due to the covid situation, the consumption is lower compared to the previous years. Nevertheless, all the peaks went forward one hour, occurring respectively from 7:00 p.m. to 8:00 p.m. in the evening and from 9 a.m. to 11:00 a.m. in the morning. This shift can highlight an issue in data pre-processing toward DST time change. After, more information from Borkum, a shift may indeed have been applied in the Excel files on several periods. Thus, in order to keep the correlation between Borkum local hours and consumption behaviour and because of the specificity of 2020 due to covid situation, the consumption load curve collected in 2020 will not be used in the historic dataset to calibrate the forecasting model.







This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

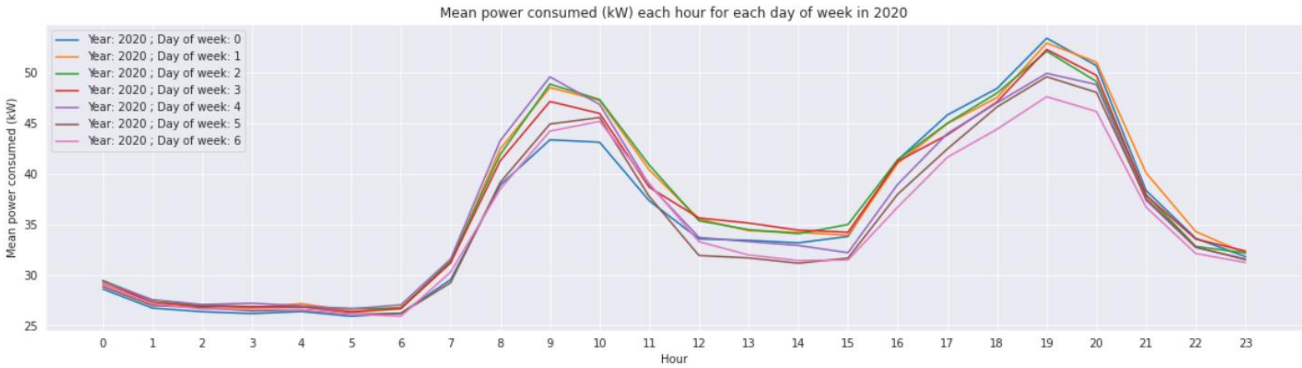


Figure 9 Mean hourly consumption during the years 2017 to 2020 expressed in kWh per hour.

The previous Figure 9 suggested a global behaviour. The variations among one year can be better analysed by plotting the boxplots of the energy consumed during each hour. For example, Figure 10 highlights that on Tuesdays (day = 1), 50 % of the time, the energy consumed from 8:00 p.m. to 9:00 p.m. (boxplot above hour = 20 in the following graph) was between 20 kW and 115 kW.

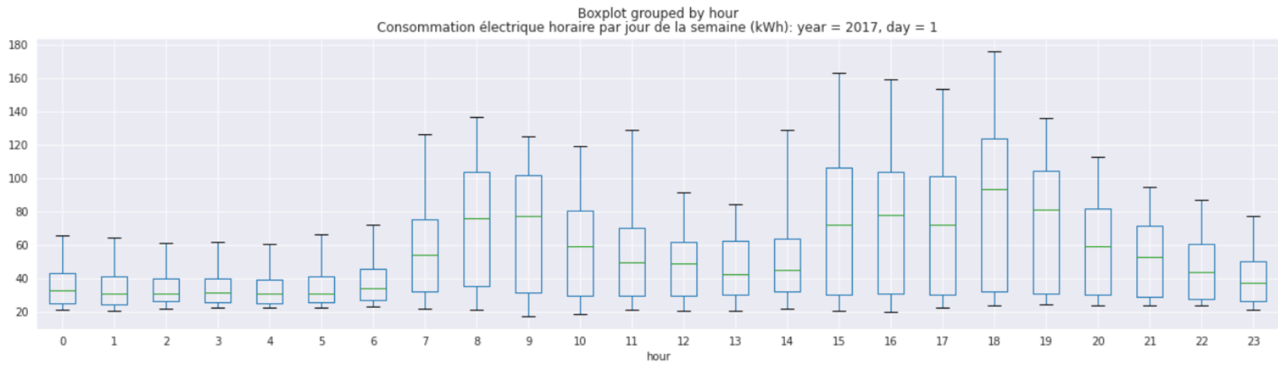


Figure 10 Histplots of hourly consumption of the year 2017.

6.2.1 Household end-users

6.2.1.1 General information on the data

Thirty households have been selected to be involved in Islander project. Their annual consumption is in between 2000 kWh and 7000 kWh. Currently and until smart meter will be installed, there is no consumption data available for these households in Borkum. The models have been tested using anonymous consumption data of BCM's clients that agreed to share their data.

Therefore, the consumption data of a French household has been collected for one full year at a granularity of 1 hour (2021) as shown in Figure 11. Very few load curves were available with a historical length of a year.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

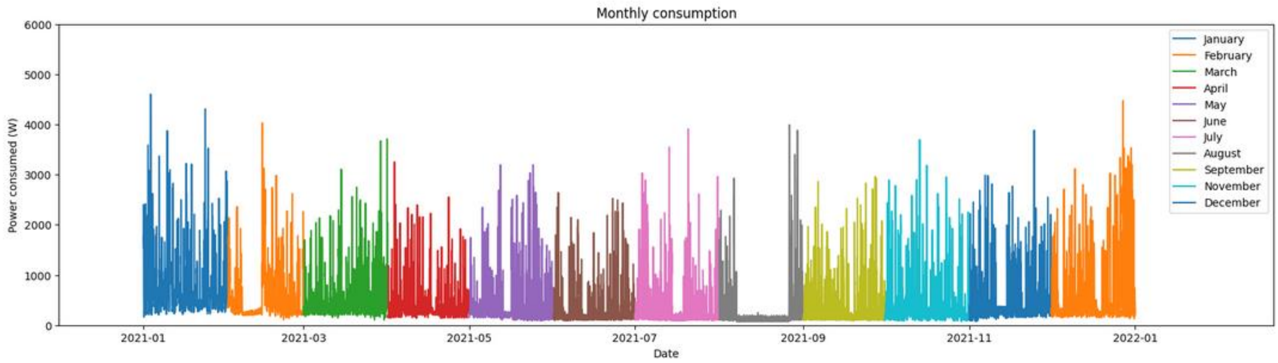


Figure 11 Household load curve on which the individual consumption model is tested (expressed in kW versus time)

6.2.1.2 Data cleaning

It is not required to perform a cleaning process because the data from Enedis are usually of high quality. The only issues encountered are detailed in Table 2.

Issue	Assumptions / actions
Missing values (19 hourly data missing in total)	Prophet library handles the missing values. It is considered in the analysis section

Table 2 Household data cleaning actions and assumptions

6.2.1.3 Data analysis

6.2.1.3.1 General analysis

All the values are over a threshold of 68 W, meaning that even if the house is unoccupied, some equipment - such as a fridge, a freezer, a TV box, a boiler or also electric heaters - are always turned on or have an auxiliary consumption. This threshold is equivalent of an annual consumption equal to 0,068 * 365*24 ≃ 596 kWh on the full year. This threshold was increased in winter (consumption of heating equipment).

The consumption profile when the house is unoccupied is plotted in Figure 12. Studying the consumption when the house is unoccupied is relevant since this noise is present all over the year. In particular, the consumption (kW) oscillated between 0.11 kW and 0.18 kW, with five lower values reaching 0.08 kW (5 hours in a month).

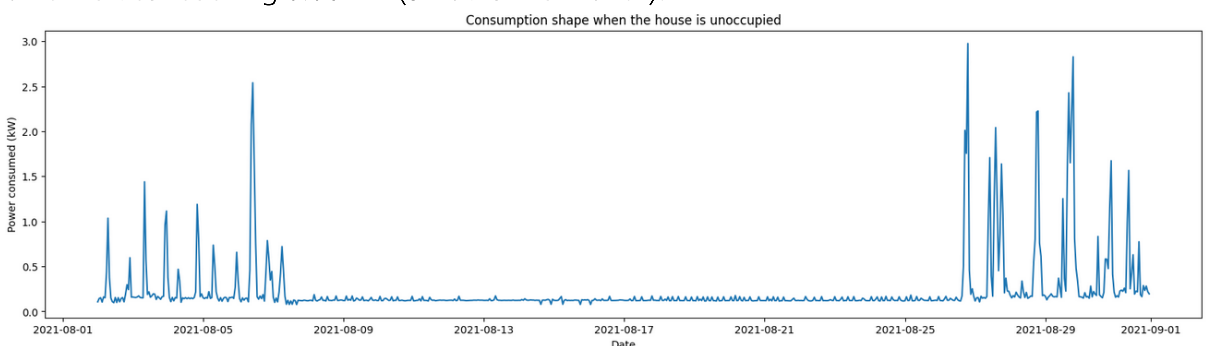
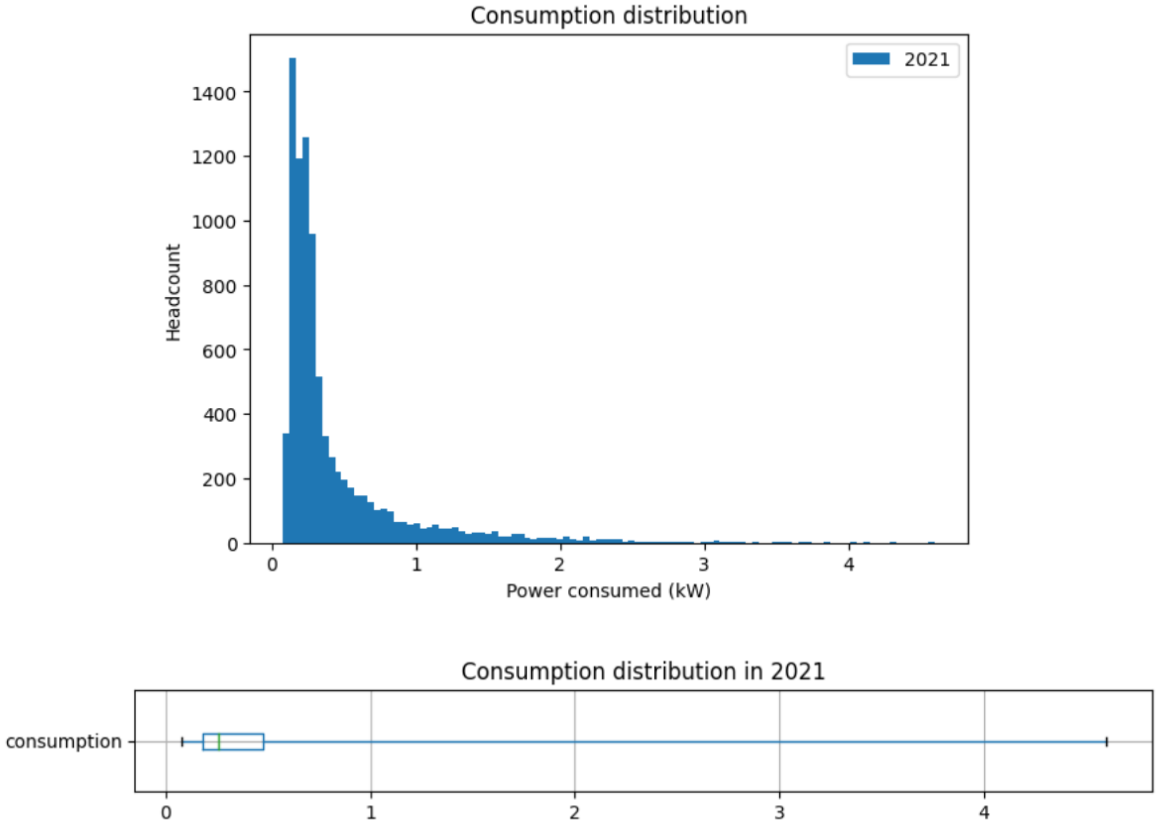


Figure 12 Consumption shape when the house is not occupied during August 2021.

The consumption distribution (Figure 13) showed that the power is lower than 0.959 kW 99%

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

of the time. The maximal power reached is 4.602 kW.




```
Consumption distribution in 2021
count    8753.000000
mean        0.439021
std         0.466178
min         0.076000
25%         0.181000
50%         0.260000
75%         0.478000
max         4.602000
Name: consumption, dtype: float64
```

Figure 13 Consumption distribution during the year 2021. The values of the table are expressed in kW.

6.2.1.3.1 Seasonal analysis

Yearly trend

Electric consumption is higher in Winter compared to Summer (probably the heating system) as shown in Figure 14.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

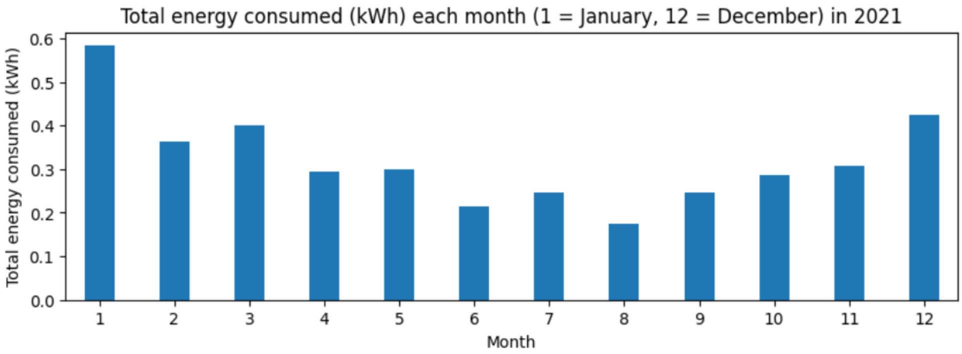


Figure 14 Monthly power consumed in 2021 (kW per month).

Weekly trend

On a finer scale, the energy consumed each day is in between 9 and 12 kWh, 12 kWh being reached on Sundays (Figure 15).

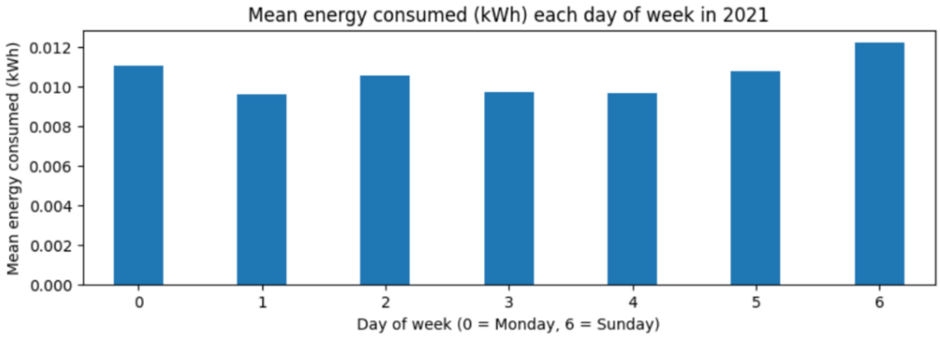
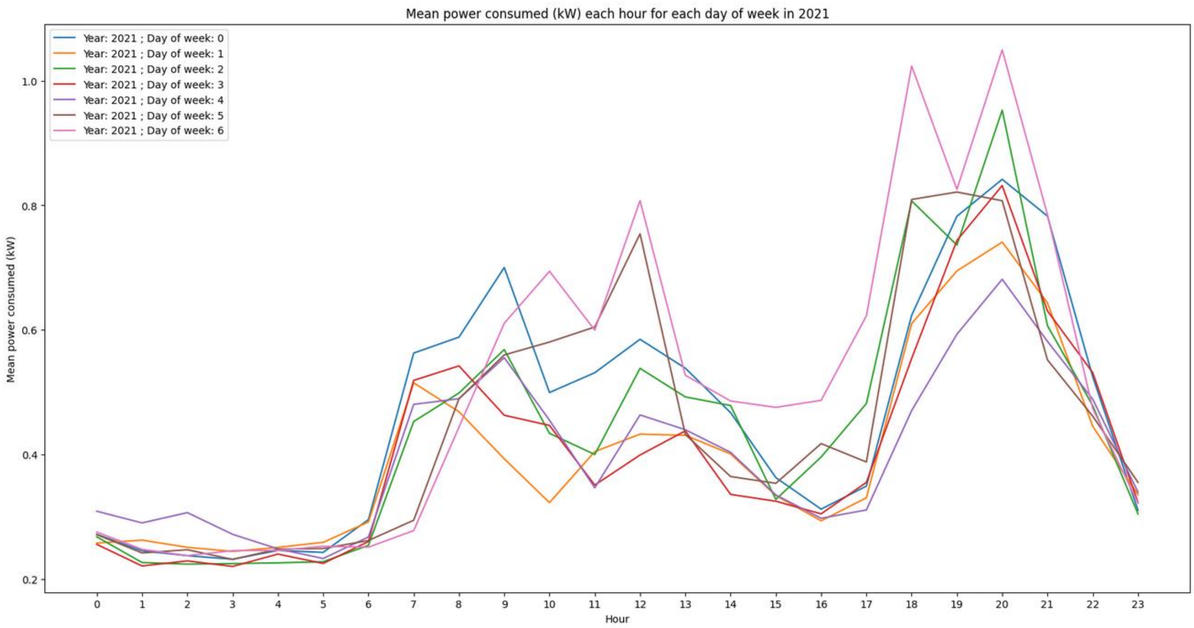


Figure 15 Mean daily power consumed in 2021 (kW per day).

Hourly trend

Figure 16 highlights the hourly consumption trend every day of the week, by computing the mean power consumed hourly from the whole year of historic data. In the evening, peaks can be noticed mostly from 6:00 p.m. to 7:00 p.m. (hour = 18 on Figure 16) and from 8:00 p.m. to 9:00 p.m. (hour = 20 on the figure). These peaks tend to be higher on Sundays (day of week = 6) and lower on Wednesdays (day of week = 4).



This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

Figure 16 Weekly shape of the power consumed in 2021 (kW per hour).

The boxplot (Figure 17) highlights that half of the time on Sundays (day of week = 6), the energy consumed from 8:00 p.m. to 9:00 p.m. (boxplot above hour = 20 in the following graph) was between 0.5 kW and 1.5 kW. The power consumed from 8:00 p.m. to 9:00 p.m. thus is not always the highest consumption of the day on Sunday.

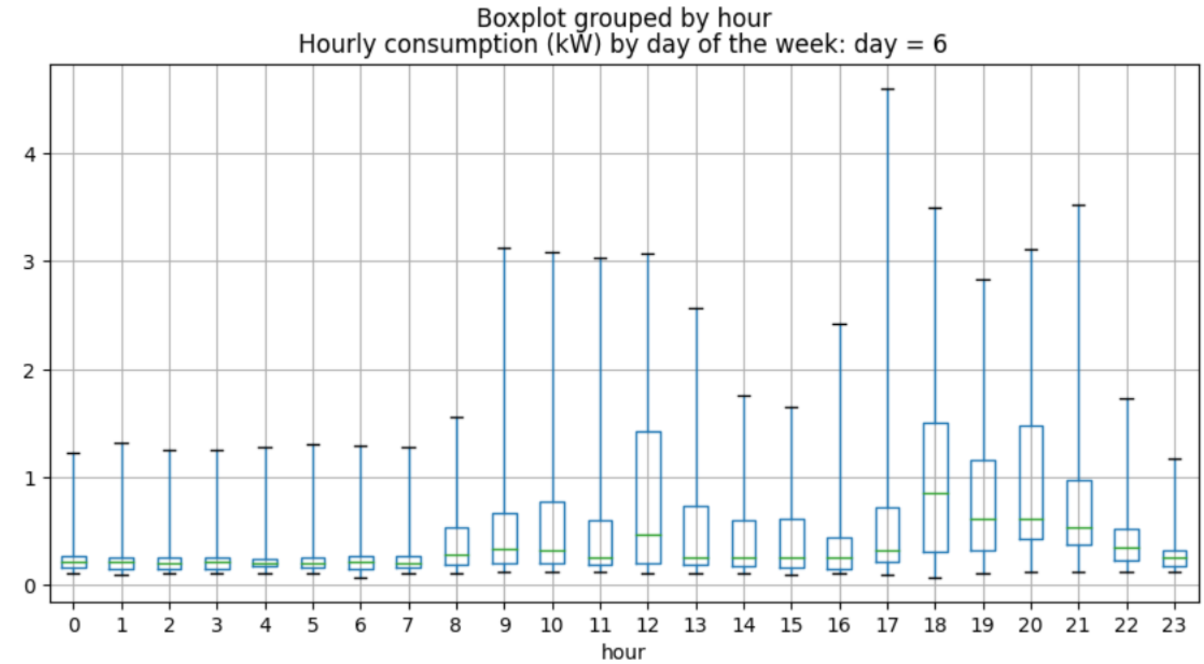


Figure 17 Boxplot of the hourly shape of the power consumed in 2021 (kW per hour). The green line in each box represents the median value of the column by default.

7 FORECAST MODULE DEVELOPMENT

7.1 Grant agreement requirements

Due to certain limitations, it was not feasible to adhere to the proposed project plan of the grant agreement. Firstly, when model development started, there was no data pertaining to the daily activities of households and businesses available. Additionally, the tariff structure on the island remains uniform for all consumers, eliminating the possibility of utilizing tariff-related data. Moreover, the unavailability of real consumption data further restricts the execution of the proposed plan, as it is not possible to perform a segmentation of the clients. Additionally, the extraction of controllable loads is hindered by the absence of smart meters to measure individual equipment consumption, which will not be considered in the project due to the costs and difficulties of implementation. Consequently, performing client segmentation and extracting controllable loads becomes impractical currently.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

7.2 Modelling strategy

The approach to implement the forecasting model was the following. The first step was to define the specifications: goal, inputs of the model, outputs of the model. The data was then collected, cleaned and analyzed (paragraph 6.2). Then, multiple model approaches have been tested for model development:

Static model: Borkum utility coefficients

Borkum utility provided seasonal coefficients to build the shape of the load curve (normalize coefficients for each 15 minutes of the year). The coefficients are linked to a profile (bakery, small business, household, etc.). The model developed takes the profile (and therefore the coefficients) and the historical consumption data (with a minimum of 1 year of historical data) as inputs. Then, the forecasted power is computed by multiplying the shape given by the coefficients with the mean annual consumption over 15 minutes (in kW per 15 minutes).

This approach has the advantage of being very simple. However, it requires that the annual consumption is available and did not provide a high accuracy.

Static model: Enedis coefficients

The second model tested was very similar to the model using Borkum utility coefficients. Indeed, it considered Enedis (the French DSO) coefficients that are more precisely computed and that are dynamical (they update the coefficients). The accuracy was higher compared to Borkum coefficients; however, it was not satisfying enough. Therefore, it has been chosen to test a dynamical solution.

Machine learning model using prophet library from Facebook

Prophet is a time series forecasting algorithm developed by Facebook's Core Data Science team. It was designed to address the challenges of forecasting data with seasonal patterns, non-linear trends, and outliers. Prophet incorporates a flexible model that captures various components of time series data, including trend, seasonality, and holiday effects. It utilizes an additive model approach, where these components are combined linearly to generate forecasts. The algorithm also incorporates a Bayesian framework for modelling uncertainty and provides intuitive parameter tuning options for users. Prophet has gained popularity due to its ease of use, robust performance, and ability to handle a wide range of time series forecasting tasks [4].

It has been chosen to forecast the individual consumption using the Prophet software because it is an efficient machine learning approach that considers all the parameters chosen which are weather conditions, real-time consumption data, etc (as shown in the grant agreement). New parameters can be added afterward, which makes it relevant for the project (as all input data are not available yet). Also, it provided interesting results to forecast the global consumption of the island of Borkum [5], and has multiple parameters and hyperparameters that can be optimized to improve model accuracy (multiple parameters can be used according to the type of profile – B2B or B2C). Model characteristics are detailed in deliverable 4.5 of the project [5].

This approach has been selected as it can be easily adaptable to multiple profiles (B2B and B2C) by changing the parameters and hyperparameters and is well known in the time-series forecasting field.

Finally, the parameters and hyperparameters giving the best accuracy were selected. It is important to insist again that there is no load curve from Borkum end-users that is available at the end of June 2023. To build a forecast and measure the accuracy, it is mandatory to have at least one year of historical data, which will not be available before the end of 2024.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

7.3 Calibration process of the dynamic model

The calibration process consists in testing a large number of combinations of hyperparameters and regressors, to select the combination for which the model is the most accurate among metrics wisely chosen (see "7.2.2.2.2 Model calibration" in deliverable 4.5 [5]).

For a given combination of hyperparameters (historic length, horizon, granularity, time zone of the data, etc), the model is launched on several publication dates spread over the period covered by the training dataset. For each batch of forecasts, the MAPE, RMSE, MAE and BIAS are calculated by crossing consumption forecasts with realized consumption values. Finally, the average of the metrics of each batch is calculated.

The accuracy of the model associated to this combination can thus be measured thanks to:

- The mean MAPE (respectively RMSE, MAE, BIAS) previously calculated.
- The series MAPE (respectively RMSE, MAE, BIAS) calculated on each batch of forecasts.
- The shape of the forecasts from each publication date d on a horizon H.

To have a statistically representative model accuracy, it is important to launch enough batches of forecasts (see 8.1.2.7. Impact of the number of simulations on the model performance analysis [5]). Testing the model consisted in running the batches of forecasts each 25 hours of at least one full year. This interval between two publication dates has been chosen so that each hour and each day of the week could be represented in the performance model analysis.

It has been chosen to calibrate the model with a horizon of 1 day because the forecasts will be updated and sent every day to the smart IT platform (with a horizon of 2 days). The accuracy will be measured with a horizon of 2 days (see 8.1.2.6. Impact of the horizon of forecasts on the model performance analysis [5]).
In order to automate and optimize the duration of the calibration process, an informatic package constituted of several modules has been developed in this task 4.1. As the load curves are available only for one year, excluding 2020, the calibration process will be done in 2019 for the youth hostel consumption forecasting model, and in 2021 for the household consumption. Many articles. suggested that the grid search method was efficient in finding the optimal parameters [6], [7]. Grid-searching aims at scanning the data to configure optimal parameters for a given model. This method has been implemented in the task 4.1.

8 RESULTS

This section is divided in 2 parts, on section dedicated to the results of the social behavioural study on the energy consumption habits of Borkum household and businesses (paragraph 8.1).
The second section is dedicated to short-term consumption forecasting (paragraph 8.26.2).

8.1 Island consumer energy habits

8.1.1 Commercial end-users

The three buildings report a different amount of monthly energy consumption. Building 1

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

reports a consumption of 500 kWh a month, building 2 reports a consumption of 31836 kWh a month, and building 3 consumes around 2400 kWh a month. The fact that building 2 has many employees active in the building might explain the higher level of energy consumption. However, as building 3 has the lowest number of employees but not the lowest energy consumption, another explanation could be possible. The three buildings report to use a different number of varying appliances (fridge, computer, …). Building 1 has 48 appliances, building 2 has 53, and building 3 has 97 appliances. Building 2 is not reporting to have the most appliances in use, however, they are reporting the largest number of "big" appliances in the building, such as computers (N = 30), on-site servers (N = 3), electronic displays (N = 5) … Given that many of those appliances are running 40 hours a week, this could explain the high energy demand from building 2. Building 3 is reporting many domestic appliances, but less "big" appliances, and building 1 is reporting the lowest number of appliances of the three buildings.

8.1.2 Household end-users

The goal of the data collection was to be able to develop an overview of different type of end-users based on their demographical characteristics and energy consumption behaviour. To do so, we conducted a "consumer segmentation" analysis using the k-means algorithm. This allowed us to define different clusters of end-users based on their different attributes. We started with the "Elbow" method to investigate the optimal number of clusters for the k-means algorithm. It is important to note that missing values were coded as 9999 in the dataset as missing values are not accepted by the algorithm. As can be seen in Figure 18, three clusters were optimal for the algorithm.

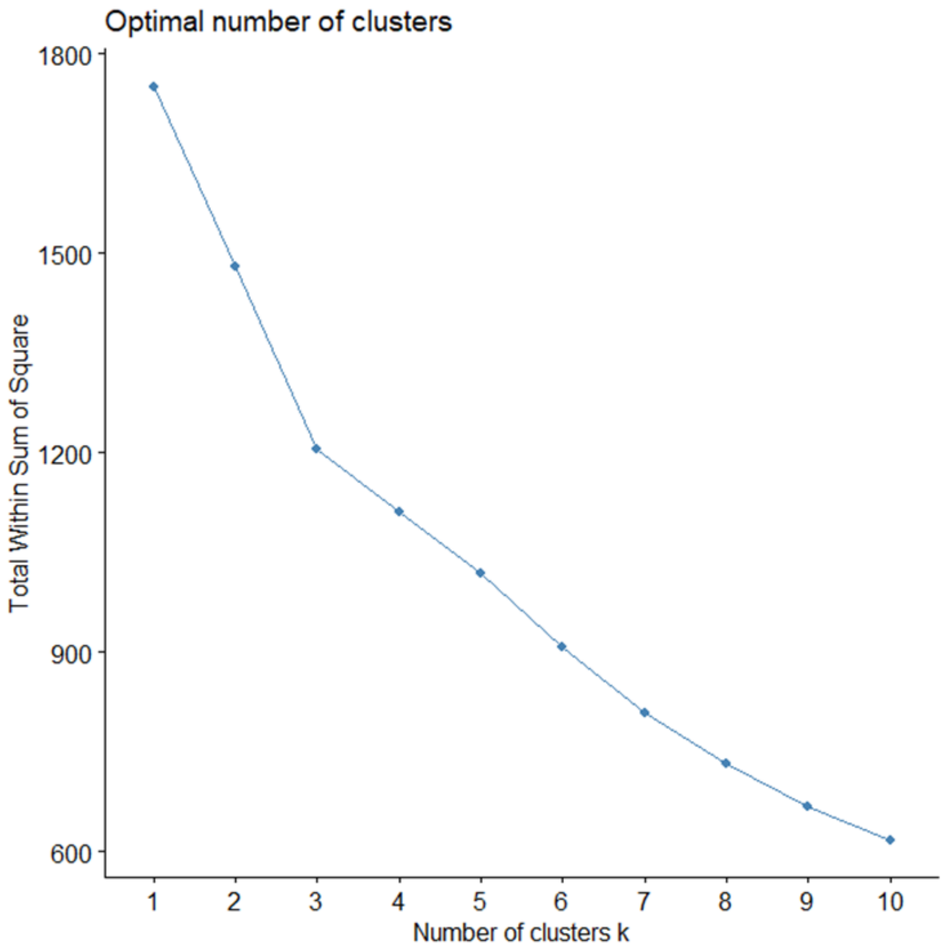


Figure 18 Optimal number of clusters in k-means algorithm.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

The results of the k-means algorithm showed three end-users clusters, with 10 end-users in cluster 1, 7 end-users in cluster 2, and 9 end-users in cluster 3. Based on the characteristics of the three clusters, we see that the difference between the three rely mostly in the mean age, household composition and time of activities. Cluster 1 consists mostly of middle-aged (50 years old) end-users living with another person, in a larger house, with many appliances, having most of their activities on Friday and Saturday. Cluster 2 consists of slightly younger (40 years old) end-users living with multiple persons, doing most of their activities on Tuesday and Friday. Cluster 3 consists of slightly older end-users (60 years old) living alone, in a smaller house with a lower number of appliances, doing most of their activities on Sunday.

However, it worth mentioning that given the fact that only two participants completed the diary survey for 7 days straight, the preliminary conclusions are based on flawed data. This means that we will need more and more complete data in the future to refine the algorithm and perfection the end-users' clusters.

8.2 Forecasting the power consumed

8.2.1 Commercial end-users

8.2.1.1 Parameters tested and results

Multiple parameters and hyperparameters were tested on the model. Only a summary of the results will be displayed in Table 3, the more detailed analysis will be written in the Annex B.

The metric used in the results is measured in MAPE of the whole year 2019.

Tested parameter	Results	Conclusions
Local time zone versus UTC time zone	UTC = 14% Local = 13%	Local time-zone will be selected
Granularity (15 minutes versus 1 hour)	15 min = 15% 1 h = 13%	1 hour granularity will be selected
Number of simulations	25-hour time lag between test batches gave better accuracy	25-hour selected
Historical length (35 days, 2 months, 3 months and 7 months)	35-days of historical length gave better accuracy	35-days selected
Temperature	No improvement of the model accuracy	This regressor will not be used for this load curve, however it will be tested on the end-user's load curve (in case they use heating or cooling electrical systems).
Holidays	No improvement of the model accuracy	This regressor hasn't been used

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

Hyperparameters of the prophet model • Changepoint prior scale • Changepoint range • Fourier order	Improvement of the model accuracy	The best parameters have been selected for the final version of the model

Table 3 Parameters tested to improve the business consumption model.

To improve the model, plots have been made to better understand which consumption behaviours are not understood by the model - with the chosen initial combination of hyperparameters, and the hypothetic causes of the forecast errors.

8.2.1.2 Microscopic forecast analysis

Specific days of the forecast have been analysed. For 08/02/2019, it seems the trend is modelled but peaks are not detected (Figure 19). This may be improved by increasing the changepoint prior scale parameter. This parameter is indeed responsible for the flexibility of the model to react quickly to a change in the trend.

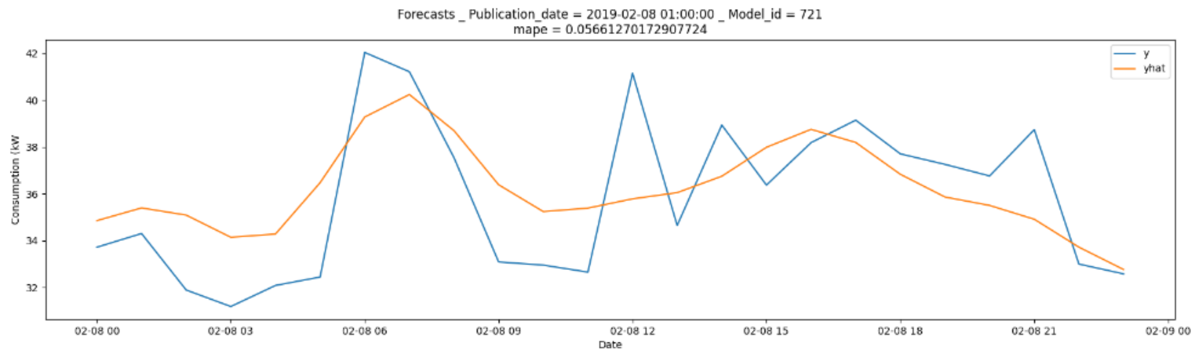


Figure 19 Consumption realized in blue versus forecasted in orange for one day (08/02/2019).

For 22/09/2019 (Figure 20), the forecast has a low accuracy due to the prediction of a consumption peak which did not occur. A decrease of consumption can be due to an event, happening that day. The holidays component of the Facebook's library lets the possibility to specify those special days so that the model would be more flexible to unexpected changes.

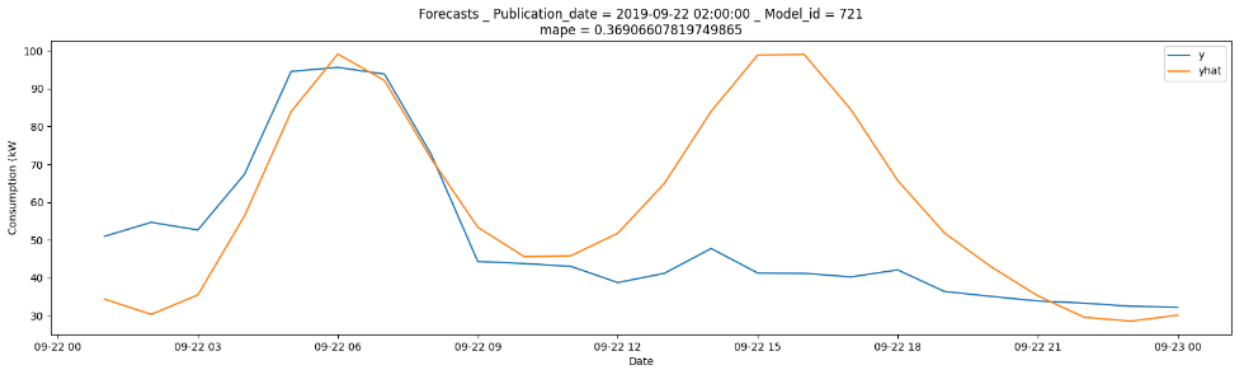


Figure 20 Consumption realized in blue versus forecasted in orange for one day (22/09/2019).

Other forecasts have a lower accuracy due to a consumption peak occurring earlier than expected, such as on the 2019-10-11 where a peak was expected around 4 p.m. but occurred

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

around 2 p.m. (Figure 21).

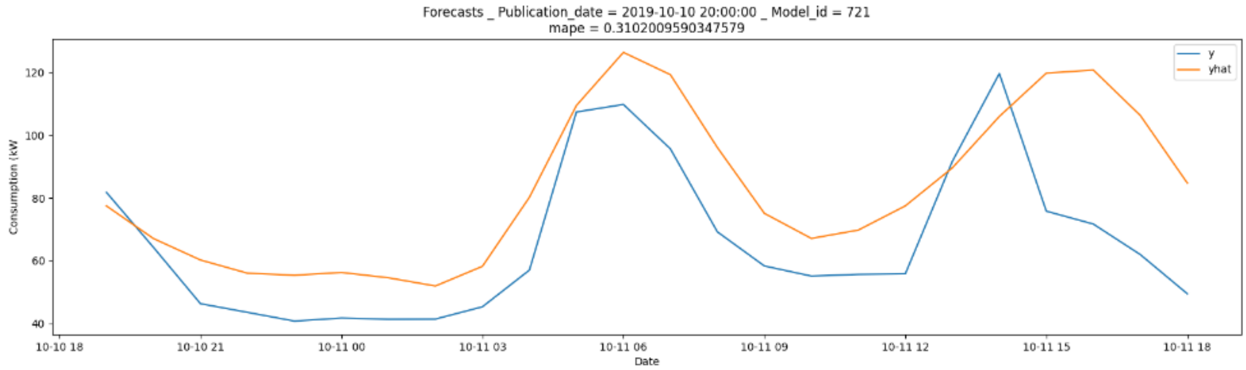


Figure 21 Consumption realized in blue versus forecasted in orange for one day (22/09/2019)

At least, the trend is often followed by the model but with a consumption volume over-estimated or under-estimated. These volume gaps can be reduced by including an additional regressor to the model like the youth hostel realised and predicted gauge. In the case of this youth hostel, data are not available. If the youth hostel gauge is correlated to the number of tourists coming on the island, adding this regressor could thus improve the forecasts.

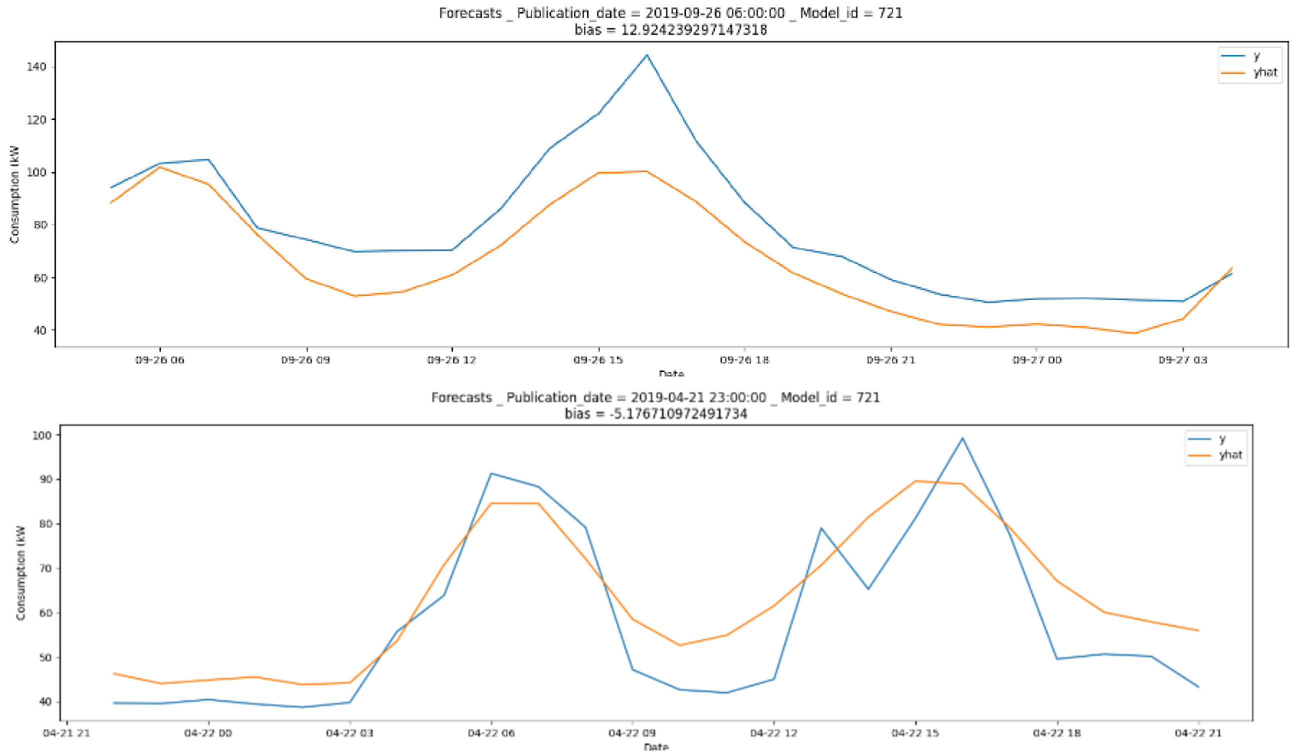


Figure 22 Consumption realized in blue versus forecasted in orange for two days (26/09/2019 & 21/04/2019).

8.2.1.3 Macroscopic forecast analysis

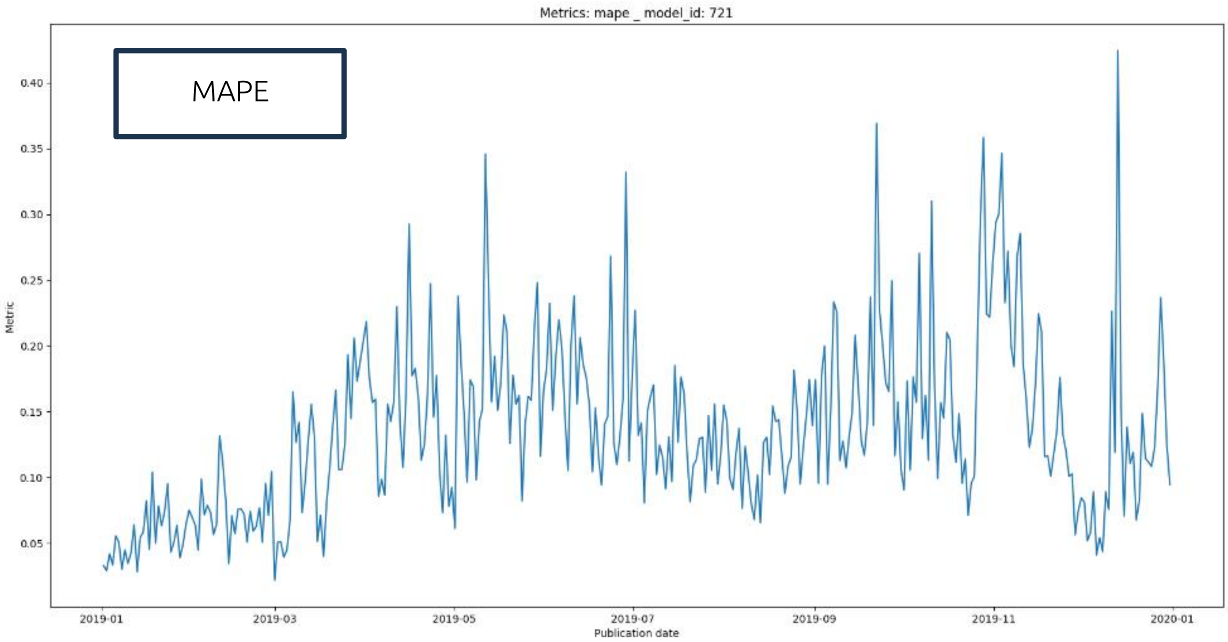
Each of the metrics MAPE, RMSE, BIAS and MAE bring information about the accuracy of the forecasts on the period where they have been calculated (formulas are written in deliverable 4.5 [5]). The MAE - Mean Absolute Error - is used to measure the absolute gap between real values and forecasts on average. An MAE equal to zero would describe a model without error on the period. The MAPE is more sensitive to the errors made during a period of low

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

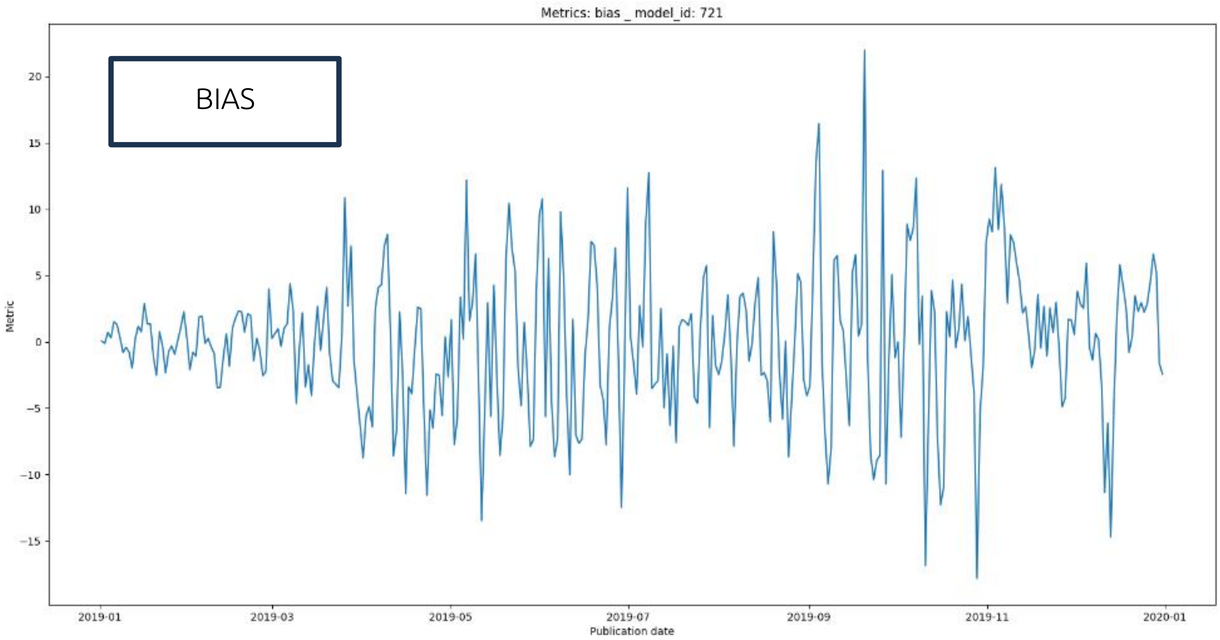
consumption. The RMSE focuses on the errors made on high consumption data, thus among other things on the forecasts of consumption peaks. The BIAS calculated for a batch of forecasts on a one-day horizon gives information about how much the consumption has been under or over-estimated. The results should nevertheless be analysed carefully since over-estimations and under-estimations can be balanced on the horizon of forecasts. Plotting the MAPE in parallel of the BIAS is thus a way to ensure that a small value of the bias is not hiding large forecast errors highlighted by a high MAPE. Table 4 and Figure 23 show the results reached.

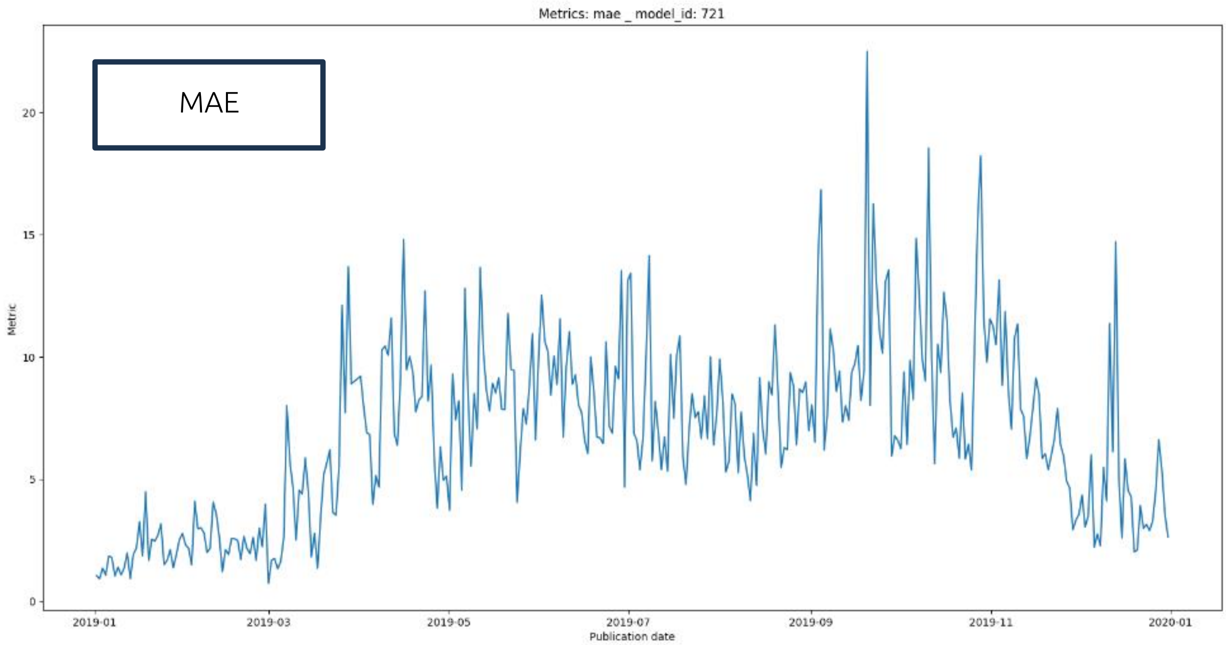
Month	Mean MAPE (%)	Mean RMSE (kW)	Mean BIAS (kW)	Mean MAE (kW)
January 2019	0.05	2.69	0.14	1.94
February 2019	0.07	3.24	0.04	2.54
March 2019	0.11	6.54	0.56	4.85
April 2019	0.15	10.16	-2.39	8.18
May 2019	0.17	10.94	-0.42	8.3
June 2019	0.17	12.2	-1.06	8.79
July 2019	0.13	10.76	-0.14	7.83
August 2019	0.12	10.04	-0.29	7.37
September 2019	0.17	13.4	0.5	10.32
October	0.17	12.23	-1.03	9.77
November 2019	0.18	8.74	3.43	7.6
December 2019	0.11	5.44	0.59	4.6
Whole year	**0.13**	**8.91**	**0.05**	**6.86**

Table 4 Results of MAPE, RMSE, BIAS and MAE for the whole year.



This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669





This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

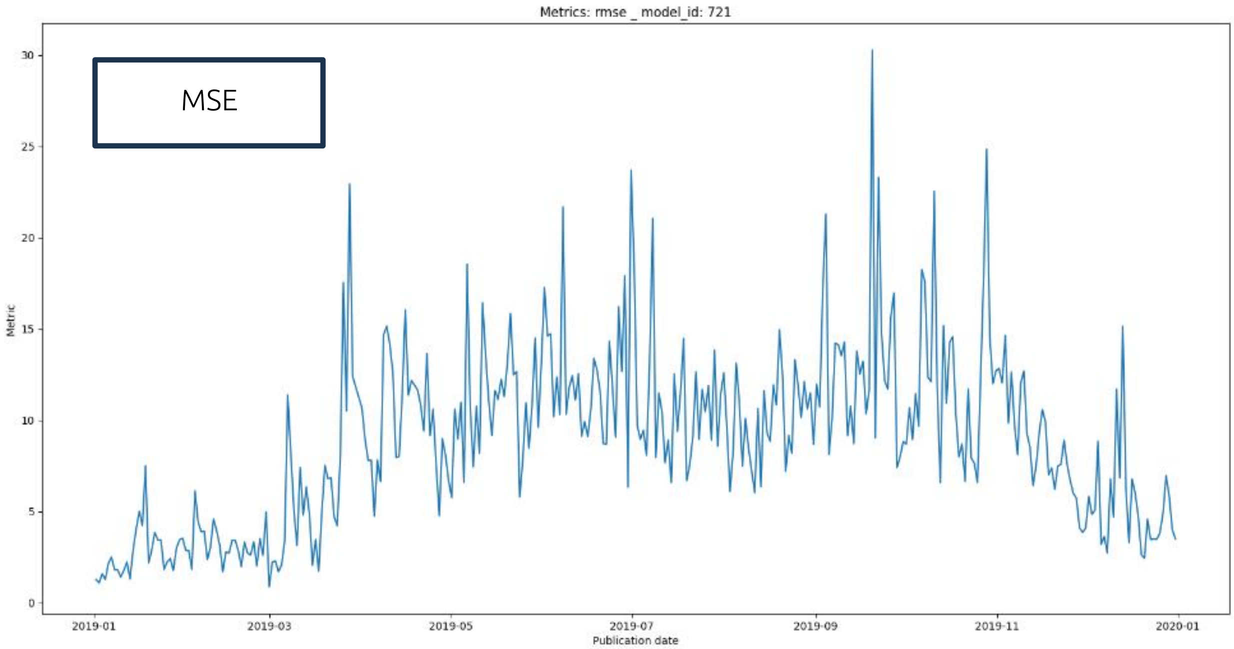


Figure 23 Consumption forecast metrics for the year 2019. MAPE, BIAS, RMSE and MSE are computed.

In December, January, February and March, consumption forecasts are very accurate with a MAPE most of the time under 20%. As Borkum is known to receive less tourists in winter, the accuracy of the forecasts can be explained by the lowest variation of consumption behaviours and less consumption peaks than in summer.

There is a decrease in the MAPE, RMSE and MAE until the end of June, with several peaks of the MAPE over 20%. This is also the case in September and in October where the highest values of the MAPE are reached. These peaks can be due to the fluctuations of the number of tourists and inhabitants on the island due to holidays or to special events and festivals.

In July and August, the MAPE is lower with no peaks.

8.2.1.4 Discussion

The results reached are encouraging. However, as no data is available from the 3 end-users selected on Borkum, more research should be done after 1 year collecting historical data. Each model should be tested on a larger number of sample of businesses with the same activity to ensure the accuracy reached is like the one tested in this document. The results of the social demographic questionnaires (paragraph 8.1) could increase forecast accuracy as it provides information on their energy equipment and source of consumption (which was not available for the load curve studied in this deliverable.

8.2.2 Household end-users

8.2.2.1 Parameters tested and results

As a first approach, the parameters of the Prophet forecasting model (explained in paragraph 7.2) used to forecast the global consumption on the island in deliverable 4.5 [5] were applied to the French load curve (model_id = 930). Indeed, individual consumption behaviours include trends and seasonality which were considered in the previous model. Nevertheless, the big difference with the aggregated consumption of the whole island is the variability of

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

consumption peaks. Indeed, even if a household has consumption habits, the occupants can move their electric usages from an hour or day to another whereas aggregating many consumption load curves has a smoothing effect on the peaks. The accuracy of the previous model is thus expected to be much lower due to the variability of the individual consumption peaks.

Multiple parameters and hyperparameters were tested on the model. Only a summary of the results will be displayed in Table 5Table 3, the more detailed analysis will be written in the Annex C.
The metric used in the results is measured in MAPE of the whole year 2021.

Tested parameter	Results	Conclusions
Local time zone versus UTC time zone	Not tested	Local time zone will be selected (results from paragraph 8.2.1.1)
Granularity (15 minutes versus 1 hour)	Not tested	1 hour granularity will be selected (results from paragraph 8.2.1.1)
Number of simulations	25-hour time lag between test batches gave better accuracy	25-hour selected (results from paragraph 8.2.1.1)
Historical length (35 days, 2 months, 3 months)	3-months of historical length gave better accuracy	3-months selected
Holidays	No improve in model accuracy noted	This regressor hasn't been used
Hyperparameters of the prophet model • Changepoint prior scale • Changepoint range • Fourier order	Improve of model accuracy	The best parameters have been selected for the final version of the model

Table 5 Parameters tested to improve the household consumption model.

8.2.2.2 Macroscopic forecast analysis

Table 6 shows the results of the model applied to the household load curve. To have data on the whole year, a historical length of 35 days was selected.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

Month	Mean MAPE (%)	Mean RMSE (kW)	Mean BIAS (kW)	Mean MAE (kW)
January 2021	/	/	/	/
February 2021	0.88	0.49	-0.03	0.37
March 2021	0.55	0.40	0.04	0.28
April 2021	0.61	0.35	0.01	0.23
May 2021	0.79	0.37	0.00	0.26
June 2021	0.53	0.26	0.03	0.16
July 2021	0.81	0.34	0.02	0.23
August 2021	0.82	0.24	0.02	0.17
September 2021	0.71	0.31	0.01	0.21
October 2021	0.74	0.36	-0.02	0.24
November 2021	0.51	0.33	0.01	0.21
December 2021	0.57	0.43	0.03	0.30
Whole year	**0.68**	**0.35**	**0.01**	**0.24**

Table 6 Results of MAPE, RMSE, BIAS and MAE for the whole year.

As expected, the metrics of errors are higher than for the Youth Hostel because of the higher variability of the household's consumption behaviours, in particular consumption peaks and departures from the house. However, some trends and seasonality can be better fitted as daily or weekly usages, for example by increasing the length of the historic dataset, so that the model learns these behaviours from more observations.

8.2.2.3 Discussion

It should be noted that the results are not representative of the end-users of Borkum, as the metrics were computed on the consumption data of a French household. However, as no data was available, and few load curves were available at BCM (with lower historical data length most of the time). The model should be tested on more load curves to measure the accuracy more precisely.

Two questionnaires have been created to understand household consumption habits and what is the consumption linked to. This data will be useful to increase model accuracy as additional inputs should increase the learning process done by the software.

9 CONCLUSIONS

The goal of the current deliverable was on the one hand to collect data on end-users' demographics, behavioural habits, current energy consumption, and past energy consumption. On the other hand, the goal was to create a model that can predict household and businesses consumption accurately.

Concerning the first goal, by matching the different data, we aimed at developing different profiles of end-users based on their characteristics and energy consumption patterns. This would allow to develop a forecasting model capable of predicting the energy consumption pattern of Borkum' habitants based on their specific end-user profile.

The behavioural analysis revealed that in general end-users consume most of their energy during the middle of the week, and more specifically during mornings and evenings. The most reported energy consuming activities were working, watching TV, and doing laundry. When computing an end-user segmentation analysis, three groups of end-users are highlighted. Group 1 consists mostly of middle-aged (50 years old) end-users living with another person, in

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

a larger house, with many appliances, having most of their activities on Friday and Saturday. The second group consists of slightly younger (40 years old) end-users living with multiple persons, doing most of their activities on Tuesdays and Fridays. The third group consists of slightly older end-users (60 years old) living alone, in a smaller house with a lower number of appliances, doing most of their activities on Sundays. However, as noted in the deliverable report, the data were not complete, and the results cannot be conclusive. It will be needed to continue data collection to confirm or refine the groups suggested here.

Concerning the second goal, no real-time consumption data is currently available on the 30th of June 2023 (as the smart meters are not installed yet, and there is not 1 year of historical data available). Provisional consumption data was collected from Borkum utility and BCM database to test the model.

After cleaning, analysing, and testing multiple approaches, a dynamical model was selected. The modelling analysis revealed that the business consumption forecast was much more accurate compared to the household consumption forecast. It should be noted that not enough data is available to have significant results. The models should be parametrized and tested again with data from the end-users of Borkum, with many load curves.

In the course of 2023, smart meters will be installed on the island. This will allow to investigate whether end-users' behavioural reports are accurate. It will also allow the collection of end-users current energy consumption data. Once the smart meter data will be collected, and after one year of data collection, it will be possible to refine the analyses and results reported in the current version of the deliverable.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

10 DEVIATIONS

Delivery of the content is in time and to full satisfaction, without any deviations to actions planned.

11 REFERENCES

[1] 'Survey of Commercial and Institutional Energy Use - 2019'. https://www.statcan.gc.ca/en/statistical-programs/instrument/5034_Q14_V3 (accessed Jun. 15, 2023).
[2] 'U.S. Energy Information Administration - EIA - Independent Statistics and Analysis'. https://www.eia.gov/consumption/residential/data/2020/ (accessed Jun. 15, 2023).
[3] 'American Time Use Survey (ATUS)'. https://www.census.gov/programs-surveys/atus.html (accessed Jun. 15, 2023).
[4] M. Krieger, 'Time Series Analysis with Facebook Prophet: How it works and How to use it', *Medium*, Feb. 02, 2022. https://towardsdatascience.com/time-series-analysis-with-facebook-prophet-how-it-works-and-how-to-use-it-f15ecf2c0e3a (accessed Jun. 28, 2023).
[5] Sauvajon, Bowden, Courtois, Laurent, and Huchet, 'Deliverable 4.5: MACROSCOPIC ENERGY SUPPLY AND DEMAND FORECASTING MODELS'. Jun. 30, 2022.
[6] 'Statistics for Machine Learning [Book]'. https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/ (accessed Jun. 14, 2023).
[7] D. Raspopov and P. Belousov, 'Development of methods and algorithms for identification of a type of electric energy consumers using artificial intelligence and machine learning models for Smart Grid Systems', *Procedia Computer Science*, vol. 169, pp. 597–605, Jan. 2020, doi: 10.1016/j.procs.2020.02.204.
[8] P. Lusis, K. R. Khalilpour, L. Andrew, and A. Liebman, 'Short-term residential load forecasting: Impact of calendar effects and forecast granularity', *Applied Energy*, vol. 205, pp. 654–669, Nov. 2017, doi: 10.1016/j.apenergy.2017.07.114.
[9] 'The path to developing a high-performance demand forecasting model', *Artefact*. https://www.artefact.com/blog/the-path-to-developing-a-high-performance-demand-forecasting-model-part-1/ (accessed Jun. 14, 2023).

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

12 ANNEX A: DETAILS ON DATA CLEANING

Lack of information on how data was measured

To build the future consumption of a business, it is firstly important to understand how the power consumed has been measured. Indeed, the value 28.26 kW collected at 01/01/2017 00:15:00 can be either:

- An instantaneous power measured at 00:15:00
- The mean power from 00:00:00 to 00:15:00 (ending convention)
- The mean power measured from 00:15:00 to 00:30:00 (starting convention)

And each of these three cases leads to a different load curve as highlighted in Figure 24.

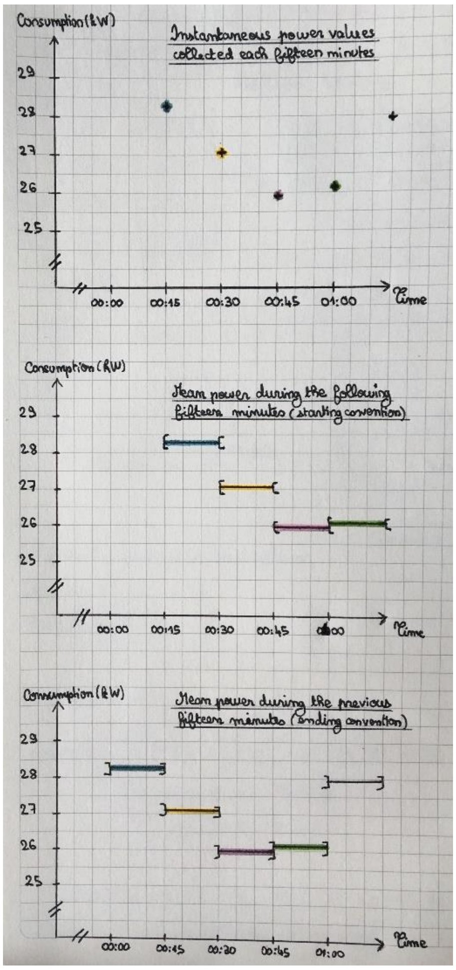


Figure 24 Schematic of the consumption considering 3 ways of measuring data.

The second and the third case bring more information since they give access to the energy consumed by the business. Indeed, the amount of energy is saved by calculating the mean of continuous power values.

As the first date of each time series is 00:15:00, it is now assumed that the power consumed at a given time step is the mean power consumed in the fifteen previous minutes (ending convention). To keep the same granularity, each time step is also rounded to the matching

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

quart of hour.

Issues on the time zone (data UTC versus local time versus a mix of them).

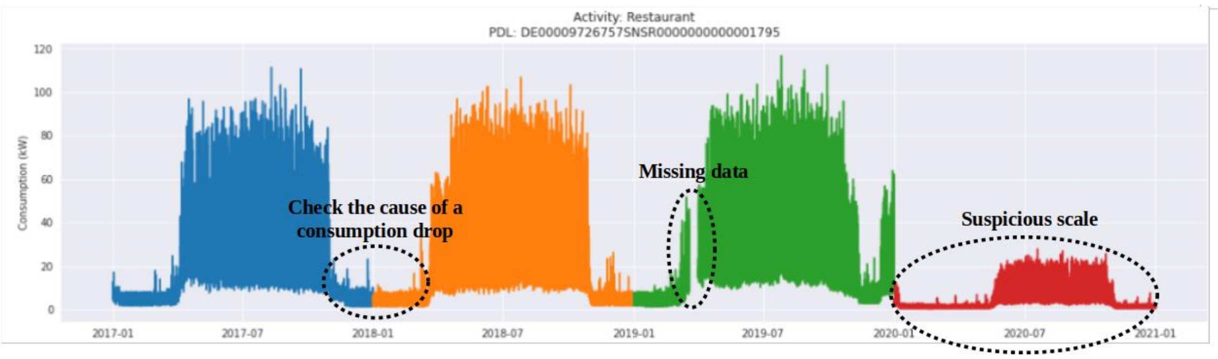
Besides, analysing a time series requires to identify the time zone in which the data have been measured: the local one – here the 'Europe/Berlin' time zone for Borkum – or the UTC time zone. If not communicated by the providers of the data, a way to check in which time zone the data have been stored is to analyse the time steps when the changing system turned to the winter or the summer hour:

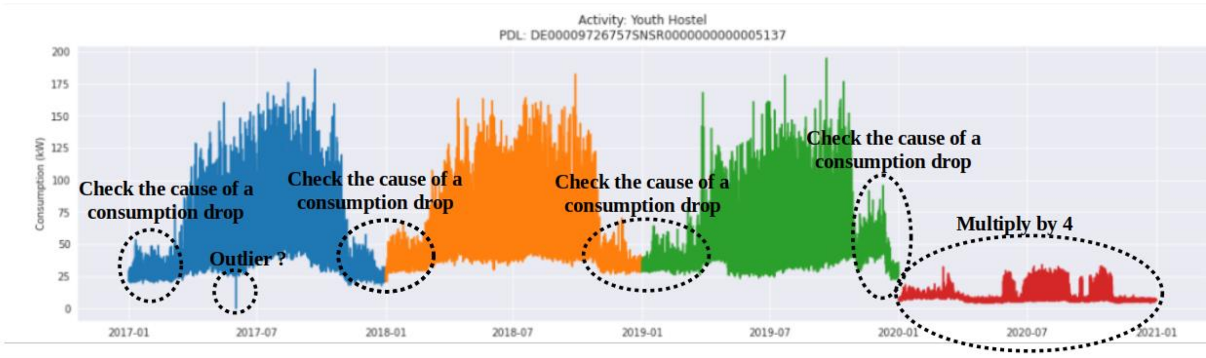
Missing time steps can be noticed at 2017-03-26, 2018-03-25 and 2019-03-31 between 2 a.m. and 3 a.m., and duplicated ones can be noticed at 2017-10-29, 2018-10-28 and 2019-10-27. These gaps and duplications highlight that the power consumed by each business in 2017, 2018 and 2019 has been collected in Borkum local time zone. On the contrary, there are no missing or duplicated time steps in 2020, which let thinks that data collected in 2020 have been stored in the UTC format.

To build the forecasting model, both UTC and Borkum local time zone should be tested into model business consumption behaviours and to predict consumption variations with the best accuracy. On the one hand, some consumption behaviour is correlated to hours in the Borkum local time zone. Indeed, Borkum's hours impact the time when people are working, eating, sleeping, etc. On the other side, consumption behaviour is often correlated to continuous phenomenon independent from the local time, such as temperature and other weather variables.

Data consistency (scale issue)

A scale issue appears for each data plotted in 2020. A division by four seems to have been applied to the data of consumption, which is not coherent with the previous years. Nevertheless, the tourist activity is known to be very low during winter months, which could also explain some of the consumption drops occurring in 2017, 2018 and 2019. Some data are also missing as shown in Figure 25.





This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

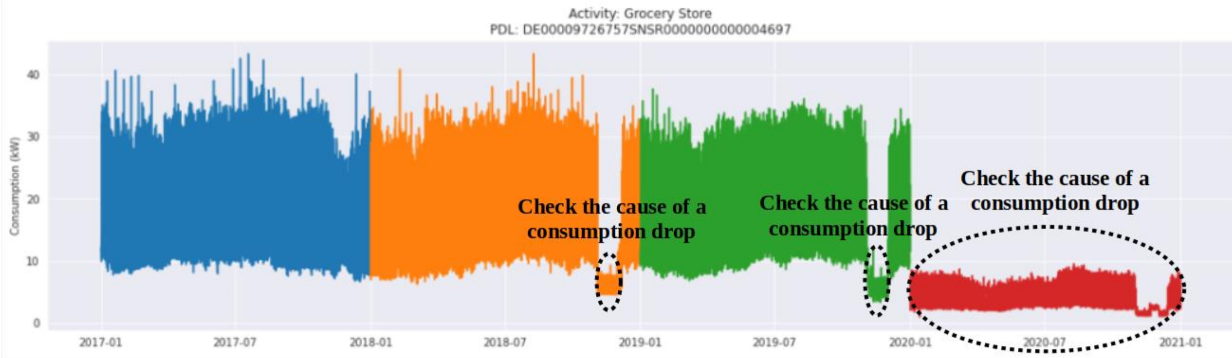


Figure 25 Load curves received by BCM before pre-treatments (consumption in kW versus time).

13 ANNEX B: DETAILS ON THE DEFINITION OF THE PARAMETERS TO FORECAST BUSINESS CONSUMPTION

13.1 Parameter 1 – Local time zone versus UTC time zone

To choose the best dates format to work with, the model was computed each day of 2019, on a horizon of one day and using an hourly consumption historic dataset of 35 days, both:

- To the consumption load curve of the youth hostel set in UTC, aggregated on an hourly granularity.
- To the consumption load curve of the youth hostel set in Borkum local time zone format, aggregated on an hourly granularity.

The calculation of the average MAPE and RMSE each month highlighted that the accuracy of the forecasts is improved in April and November, when using the local time zone (from MAPE = 16% for the UTC batch in April versus MAPE = 15% with local zone).

Finally, when comparing the whole year, the MAPE found in equal to 13% with the local time zone and 14% with UTC.

Month	Mean MAPE _ UTC	Mean MAPE _ Local	Mean RMSE _ UTC	Mean RMSE _ Local
January 2019	0.05	0.05	2.69	2.69
February 2019	0.07	0.07	3.25	3.24
March 2019	0.11	0.11	6.57	6.54
April 2019	0.16	0.15	10.53	10.16
May 2019	0.17	0.17	10.93	10.94
June 2019	0.17	0.17	12.2	12.2
July 2019	0.13	0.13	10.75	10.76
August 2019	0.12	0.12	10.03	10.04
September 2019	0.17	0.17	13.42	13.40
October 2019	0.17	0.17	12.49	12.23
November 2019	0.19	0.18	9.36	8.74
December 2019	0.12	0.12	5.42	5.44
Whole year	**0.14**	**0.13**	9.02	8.91

Table 7 Forecast results after comparing local time zone versus UTC data input.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

13.2 Parameter 2 – Granularity (15 minutes versus 1 hour)

When implementing a forecasting model on a time series, the granularity of the data is a truly important parameter to calibrate. In one hand, the smaller is the granularity, the closer are the inputs of the model to the real consumption patterns. In the other hand, a small granularity can also bring noise in the model, due to the variability of the consumption behaviours, in particular peaks. Aggregating the data at a coarser granularity allow to smooth load curve fluctuations [8], [9].

To best measure how the granularity affects the forecast accuracy, the model has thus been applied to the youth hostel's load curve at a 15-minute granularity and at an hour granularity.

The results highlight lower MAPE and RMSE when the load curve has previously been aggregated, which indicate a better accuracy of the forecasts and a best prediction of consumption peaks due to the smoothing effect. As a result, the MAPE found in equal to 15% with 15-minutes granularity and 13% with 1-hour granularity.

Month	Mean MAPE _ 15min	Mean MAPE _ 1h	Mean RMSE _ 15min	Mean RMSE _ 1h
January 2019	0.06	0.05	3.2	2.69
February 2019	0.08	0.07	3.66	3.24
March 2019	0.13	0.11	7.56	6.54
April 2019	0.16	0.15	11.47	10.16
May 2019	0.19	0.17	12.65	10.94
June 2019	0.20	0.17	14.09	12.2
July 2019	0.15	0.13	13.46	10.76
August 2019	0.14	0.12	12.12	10.04
September 2019	0.19	0.17	15.24	13.4
October 2019	0.17	0.17	13.34	12.23
November 2019	0.18	0.18	8.99	8.74
December 2019	0.12	0.11	5.6	5.44
Whole year	**0.15**	**0.13**	10.18	8.91

Table 8 Forecast results after comparing 1hour granularity and 15 minutes granularity.

13.3 Parameter 3 – Number of simulations

Calculating the mean metric for a combination of hyperparameters from 365 batches of forecasts takes approximately 10 minutes. As the Prophet model counts about 30 hyperparameters to tune, it appears thus truly important to minimize the number of simulations. However, as explain in paragraph 7.3, this number must not be too small, so that the metrics calculated be statistically representative of the model accuracy. Let's see the impact of this number of simulations on the MAPE and the RMSE.

To this purpose forecasts have been launched by batch each 25 hours and each 4 days and 1 hour. A better accuracy was found for the 25 hours tests; therefore, it was decided to test all the forecasts this way.

13.4 Parameter 4 – Impact of the length of the historical dataset

As the Prophet model learns trends and seasonality from an historic of realised consumption, the length of the historic dataset is expected to have a big impact on the model accuracy.

Four historical lengths were tested: 35 days, 2 months, 3 months, and 7 months.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

Month	Mean MAPE (%) 840 hours (model_id = 721) 1448 hours (model_id = 940) 2232 hours (model_id = 941 5208 hours (model_id = 942)			
January 2019	0.05	0.05	0.11	/
February 2019	0.07	0.07	0.07	/
March 2019	0.11	0.11	0.11	/
April 2019	0.15	0.17	0.19	0,27
May 2019	0.17	0.18	0.20	0,29
June 2019	0.17	0.17	0.19	0,25
July 2019	0.13	0.13	0.13	0,19
August 2019	0.12	0.12	0.12	0,17
September 2019	0.17	0.17	0.16	0,16
October 2019	0.17	0.17	0.15	0,25
November 2019	0.18	0.23	0.26	0,30
December 2019	0.12	0.18	0.25	0,35

Table 9 Forecast results after comparing the historical length of the input data (35 days – black values, 2 months – red values, 3 months- blue values, and 7 months – green values).

The results highlight that increasing the length of historic decreased the accuracy of the forecasts. This phenomenon can be explained by a change in the youth hostel consumption's behaviour through the different seasons in the year.

13.5 Parameter 5 - Impact of the temperature on the forecasts

When the temperature decreases or increases, heating and cooling equipment are generally used, leading to an increase of consumption, particularly in hostels whose priority is the comfort of their guests. If the youth hostel uses electric heat or cooling, the consumption could thus be highly correlated to the temperature in winter and summer.

When heat is no more useful in the rest of the year, the consumption may even be correlated to temperature, since the weather impacts the number of tourists. The results highlight no improvement of the MAPE for this parameter.

13.6 Parameter 6 - Impact of the holidays on the forecasts

Holidays can affect consumption forecasts when they are included both in the historic dataset or in the forecasting dataset of the model. The Facebook prophet library lets the possibility to specify the main holidays through a special component among the country specified, thanks to the function prophet.make_holidays.make_holidays_df.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

	ds	holiday
0	2019-01-01	Neujahr
1	2019-04-19	Karfreitag
2	2019-04-22	Ostermontag
3	2019-05-01	Erster Mai
4	2019-05-30	Christi Himmelfahrt
5	2019-06-10	Pfingstmontag
6	2019-10-03	Tag der Deutschen Einheit
7	2019-12-25	Erster Weihnachtstag
8	2019-12-26	Zweiter Weihnachtstag

Table 10 German holidays

However, it can be noticed that more German holidays may impact the youth hostel consumption. Holidays are shifted among multiple regions. The dates of national holidays or holidays of regions closer to Borkum may be more susceptible to impact the number of tourists on the island, and therefore the consumption of the youth hostel[1].

As a first analysis, only the dates suggested by the Facebook's prophet library have been specified, and a batch of forecasts has been launched each 25 hours. For some forecasts, the energy consumed is best predicted when specifying holidays.

This is for example the case when forecasting the youth hostel's consumption on 2019-12-30. It could be supposed that this improvement is due to the proximity with the Erster and Zweiter Weihnachtstag holidays, which occurred on 2019-12-25 and 2019-12-26.

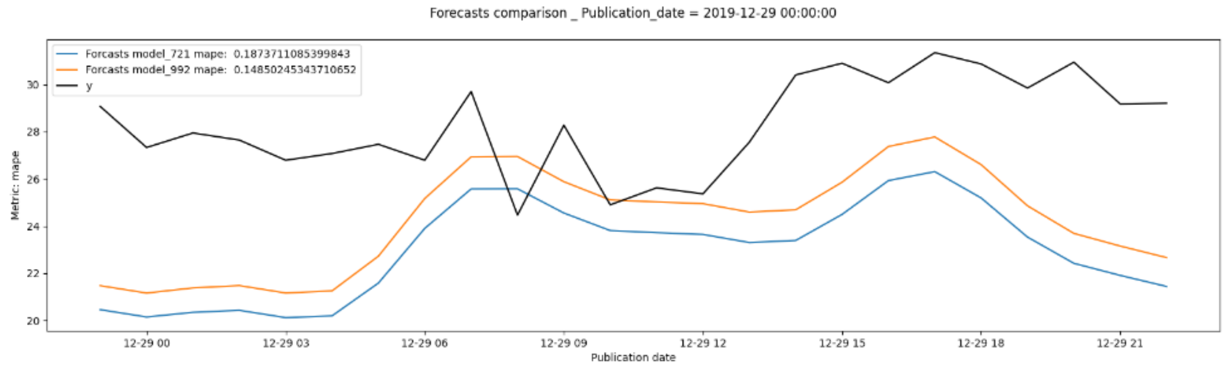


Figure 26 Comparison of the forecast accuracy for the 30/12/2019. The blue line refers to the model that does not consider the holidays; the orange to the one that considers the holidays; and the black one in the observational data.

On a global averaging scale, the specification of the holidays has not improved the model accuracy, as highlighted by the mean values of the MAPE and RMSE of Table 11. It is however possible to include more holidays in the model and even to specify windows larger than one day around a holiday to have a larger impact on the model.

[1] https://www.officeholidays.com/countries/germany/2019

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

Month	Mean MAPE (%) with holidays specification (model_id = 721)	Mean MAPE (%) without holidays specification (model_id = 922)	Mean RMSE (kW) with holidays specification (model_id = 721)	Mean RMSE (kW) without holidays specification (model_id = 922)
January 2019	0.05	0.05	2.69	2.70
February 2019	0.07	0.07	3.24	3.24
March 2019	0.11	0.11	6.54	6.53
April 2019	0.15	0.15	10.16	10.25
May 2019	0.17	0.17	10.94	10.94
June 2019	0.17	0.17	12.20	12.3
July 2019	0.13	0.13	10.76	10.76
August 2019	0.12	0.12	10.04	10.03
September 2019	0.17	0.17	13.40	13.42
October	0.17	0.17	12.23	12.18
November 2019	0.18	0.18	8.74	8.78
December 2019	0.11	0.11	5.44	5.44
Whole year	**0.13**	**0.13**	**8.91**	**8.93**

Table 11 Forecast results after comparing the if the model received as input data the holidays or not.

14 ANNEX C: DETAILS ON HOUSEHOLD MODEL PARAMETER DEFINITION

14.1 Parameter 1 – Impact of the length of the historical dataset

The following historical length was tested: 35 days, 2 months, and 3 months. The results are shown in Figure 27.

Month	Mean MAPE (%) 35 days 93 days 217 days	Mean RMSE (kW) 35 days 93 days 217 days	Mean BIAS (kW) 35 days 93 days 217 days	Mean MAE (kW) 35 days 93 days 217 days
January 2021	/ / /	/ / /	/ / /	/ / /
February 2021	0.88 / /	0.49 / /	-0.03 / /	0.37 / /
March 2021	0.55 / /	0.40 / /	0.04 / /	0.28 / /
April 2021	0.61 0.61 /	0.35 0.35 /	0.01 -0.01 /	0.23 0.22 /
May 2021	0.79 0.66 /	0.37 0.35 /	0.00 0.02 /	0.26 0.24 /
June 2021	0.53 0.56 /	0.26 0.25 /	0.03 0.02 /	0.16 0.16 /
July 2021	0.81 0.88 /	0.34 0.33 /	0.02 0.00 /	0.23 0.23 /
August 2021	0.82 0.73 0.68	0.24 0.24 0.26	0.02 0.03 0.03	0.17 0.16 0.16
September 2021	0.71 0.82 0.56	0.31 0.31 0.28	0.01 -0.05 0.03	0.21 0.22 0.18
October 2021	0.74 0.74 0.78	0.36 0.35 0.35	-0.02 -0.03 -0.05	0.24 0.24 0.24
November 2021	0.51 0.44 0.46	0.33 0.31 0.31	0.01 0.04 0.03	0.21 0.19 0.19
December 2021	0.57 0.57 0.45	0.43 0.45 0.45	0.03 0.06 0.08	0.30 0.30 0.30
From August to December 2021	**0.67 0.66 0.57**	**0.33 0.33 0.33**	**0.01 0.01 0.02**	**0.23 0.22 0.21**

Figure 27 Historical length tested on the household load curve.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

Contrary to the impact observed on consumption forecasts made earlier for the youth hostel (Annex B), the French household's consumption is more accurate when a deeper historic dataset is used. The household consumption is indeed maybe less impacted by the change of season than a youth hostel. This assumption must be tested on the load curves of the end-users of Borkum as it is not significant to test only on one load curve.

14.2 Parameter 2 - Impact of the holidays on the forecasts

To better see the impact of specifying holidays on the forecasts' accuracy, the 10 days of French holidays have been included in the model with two different values of prior scale. Zooming on the series of MAPE, the impact seems to depend on the holidays: the values obtained in May, the specification of the "Lundi de Pentecôte" has not directly improve the forecasts). On the contrary, the forecasts made after the 25 May are better after having specified the holidays. Nevertheless, these results should be carefully interpreted since they are given with a high MAPE. When zooming on the forecasts made from the publication date 2021-05-30 14:00:00, the MAPE is for example lower when specifying holidays, but the shape is as accurate computed as for the initial model. Therefore, with a global view, the forecasts were not more accurate using this regressor.

15 ANNEX D: BEST MODEL PARAMETERS SELECTED

15.1 Commercial end-users

Regarding the previous results, the best model kept has the following characteristics:

- Time zone: Borkum's time zone ('Europe/Berlin').
- Granularity: 1 hour.
- Length of historic: 35 days.
- Horizon of forecasts: 2 days.
- Regressors: consumption (kW).
- Model_id: 998.

Characteristics of the consumption load curve used for the calibration process:

- Business activity: Sports and Youth Guest House (youth hostel with accommodation, kitchen, sports centre and pitches.)
- Business localisation: Borkum (Germany)
- Metadata stored in the API Islander alpha:
    - padt= "DE00009726757SNSR0000000000005137".
    - plant= "B2B_borkum_youth_hostel_hourly_aggregated".
    - country= "DE".
    - perimeter= "B2B".
    - modelSource= "BCM".
    - meteringType= "RLM".

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

- unit= "kW".
- isForecast= False.
- isProduction=False.
- isDeprecated=False.

- Source id of the load curve stored in the API Islander alpha: a9e64765-8ff8-4bfe-b3c8-52af47bc1098.
- Year of calibration: 2019.
- Number of batches of forecasts used to calculate the metrics: one batch each 25 hours, namely 350 simulations in 365 days, thus 365 values of the MAPE, the RMSE, the BIAS and the MAE for each combination of hyperparameters tested, which have then been averaged among months and on the whole year.
- Main pre-treatment applied on the load curve on 2019: transformation into local German time zone and aggregation to a one-hour granularity.

15.2 Household end-users

Regarding the results, the best model kept has the following characteristics:

- Time zone: France's time zone ('Europe/Berlin').
- Granularity: 1 hour.
- Length of historic: 35 days.
- Horizon of forecasts: 2 days.
- Regressors: only the consumption (kW).
  Temperature in Borkum has to be tested on Borkum's household consumption load curves when they will be available.
- Model_id: 999.

The characteristics of the consumption load curve used for the calibration process are:

- Household Localisation: France
- Metadata stored in the API Islander alpha:
  - padt= "19409117194565".
  - plant= "B2C_individual_19409117194565_hourly_granularity".
  - country= "FR".
  - perimeter= "B2C".
  - modelSource= "BCM".
  - meteringType= "RLM".
  - unit= "kW".
  - isForecast= False.
  - isProduction=False.
  - isDeprecated=False.
- Source id of the load curve stored in the API Islander alpha: df330e23-0311-4a11-ad89-256cdddca88c.
- Year of calibration: 2021.
- Number of batches of forecasts used to calculate the metrics: one batch each 25 hours, namely 350 simulations in 365 days, thus 365 values of the MAPE, the RMSE, the BIAS

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669

and the MAE for each combination of hyperparameters tested, which have then been averaged among months and on the whole year.

- Main pre-treatments applied on the load curve on 2021: transformation into local French time zone and aggregation to a one-hour granularity.

This project has received funding from the European Union's Horizon 2020 research and innovation under grant agreement No. 957669