SustData: A Public Dataset for ICT4S Electric Energy Research

Lucas Pereira, Filipe Quintal, Rodolfo Gonçalves and Nuno Jardim Nunes
Madeira Interactive Technologies Institute
University of Madeira
Funchal, Madeira Islands - Portugal
{lucas.pereira, filipe.quintal, rodolfo.goncalves}@m-iti.org, njn@uma.pt

***Abstract*—Energy and environmental sustainability can benefit a lot from advances in data mining and machine learning techniques. However, those advances rely on the availability of relevant datasets required to develop, improve and validate new techniques. Only recently the first datasets were made publicly available for the energy and sustainability research community. In this paper we present a freely available dataset containing power usage and related information from 50 homes. Here we describe our dataset, the hardware and software setups used when collecting the data and how others can access it. We then discuss potential uses of this data in the future of energy eco-feedback and demand side management research.**

***Index Terms*—Public dataset, sustainability, electric energy, feedback.**

I. Introduction

Electricity consumption is steadily increasing since the 1990s, lately emerging as the second most used source of energy with a share of 17,7%, only behind oil with 40,8% [1]. One of the leading factors for this growth in electricity demand is the change in the habits of energy consumption in domestic environments. In 2010 domestic consumption was responsible for 28% of the final electricity consumption among all sectors with an effective increase of 40% between 1990 and 2010. With a 2,6% projected annual average increase it will reach 32% of the final energy consumption in the world by 2040 [2]. This surge in electricity consumption has been accompanied by an equally steady increase in the carbon dioxide emissions from fuel combustion (gas, coal and oil) used to generate electricity. These are predicted to grow by 46% until 2040 [2]. It is therefore expected that residential energy consumption will have an increasing negative impact in our eco-system, and while these impacts are hard to assess on the long-term, it is not difficult to understand the urge to reduce domestic electricity consumption. In this context households are of key importance as suggested by the literature in environmental psychology [3] and eco-feedback [4].

The work presented in this paper emerges from a large multi-disciplinary research project (SINAIS[1] – Sustainable Interactions with Social Networks, context Awareness and Innovative Services) looking at how social networks and context-awareness can be used to promote sustainable behaviors. The research team involving civil and electrical engineers, computer scientists, psychologists and designers was responsible for designing, implementing and deploying several low-cost non-intrusive electric energy monitoring systems [5]. These systems were sensing electrical circuits in many households with the goal of providing eco-feedback to the families. The research involved several deployments of different eco-feedback systems, including both qualitative and quantitative evaluation of the user interactions with the deployed systems. The overall goal was to raise the understanding and the awareness towards motivating people to consume more sustainably.

During this period of almost five years we have collected and stored large quantities of valuable data, including energy consumption, user interactions with the eco-feedback systems and electricity production from renewable sources, which we are now making publicly available to other researchers in the Information and Communication Technologies for Sustainability (ICT4S) community.

A. Related Work

In many domains, like face recognition and natural language processing, the publicly availability of datasets was fundamental in improving machine learning and data mining techniques. Currently the fields of energy, environment and sustainability research are also experiencing the emergence of publicly available datasets. For instance, just recently, the Indraprastha Institute of Information Technology released their dataset [6] containing aggregate and sub-metered electricity and gas measurements from one Indian house for the period of 73 days. This follows on others like the CASAS dataset [7] released in 2009, consisting of three months of data collected from a three-bedroom apartment located at the Washington State University. CASAS includes, among other sensor data tailored at smart-environments research, readings for temperature, hot and cold-water usage and hourly readings of electricity consumption.

Another example of an ICT4S public dataset is the Individual Household electric Power Consumption dataset [8] which is to date the longest in terms of measurement duration with about 4 years of data. This dataset was released in 2012 and comprises one-minute average measurements of the whole-

[1] http://sinais.m-iti.org

house energy consumption and the individual consumption of three sub-metered circuits, from a single house in France.

The Smart* Home Data Set [9] also released in 2012 contains a wide variety of data from three heavily instrumented houses. This dataset includes measurements from both energy usage and generation (e.g. solar and wind) and environmental data (e.g. outside and inside temperature). In addition, the authors also made available 24 hours of minute-level electricity data from 443 anonymous homes in the United States.

Within the ICT4S related datasets the Non-Intrusive Load Monitoring (NILM) community is particularly prominent given the need for extensive use of machine learning and data mining techniques. Research in this field aims at disaggregating and estimating the consumption of individual appliances by means of applying machine learning techniques to the aggregated consumption signals. NILM public datasets are expected to help researchers create more systematic evaluation processes that can be used across the different existing approaches. NILM is a very specific problem requiring the public dataset to include not only the measurements for the whole house consumption, but also information about the individual loads consumption, i.e. ground-truth data, that can be used to evaluate the performance of the different algorithms.

To date there are, to the best of our knowledge, five public datasets specifically created for NILM research, namely, the Reference Energy Disaggregation Dataset (REDD) [10], the Building-Level fUlly-labeled dataset for Electricity Disaggregation (BLUED) [11], the Almanac of Minutely Power dataset (AMPds) [12] the Pecan Street Research Institute (PSRI) [13] dataset and lastly the "UK-Dale" [14]. All these datasets provide whole house consumption data and ground-truth information of individual appliances either by labeling the power changes in the whole house consumption signal (the case of BLUED) or by providing the aggregated consumption of individual appliances or circuits in the house (REDD, AMPds, PSRI and "UK-Dale"). Furthermore, two of these datasets supplement the energy consumption data with other measurements, namely the gas and water consumption in the AMPds and photovoltaic generation in the Pecan Street dataset. These are expected to play important roles in the creation of new disaggregation strategies that include data from different sensors, i.e. sensor fusion.

Motivated by the increasing importance of public datasets for ICT4S research we decided to make our dataset public. Our data is unique because it combines both NILM specific information with long-term consumption and eco-feedback information thus providing a perfect opportunity to test NILM approaches that impact long-term studies of households and eco-feedback. We believe that SustData can also play an important role in the fields of energy, environment and sustainability. Moreover, given the nature of our data, we argue that it can be a very important contribution to the field of energy eco-feedback, enabling researchers to test different feedback techniques using fine-grained consumption data. It can also play a role in emerging fields like smart grids where accurate energy demand models are crucial for planning electricity distribution networks and optimal production capacity.

The remaining of this document is organized as follows. In section II we provide a brief presentation of the four energy monitoring and feedback deployments as well as the monitoring frameworks that were created to support them. In sections III and IV we provide in-depth descriptions of all the data items that compose our dataset. In section V we show how other researchers can access the data, which can be done either online through a custom built Application Programming Interface (API), or offline after exporting the data using the dedicated features of the mentioned API. In section VI we present some potential uses of this dataset, that we hope can help the reader to better understand its relevance to the field of ICT4S, as well as open new research perspectives. We then conclude and outline our future work plans regarding public datasets for ICT4S.

II. Energy Monitoring and Eco-feedback Deployments

The first deployment happened in July of 2010, with 17 apartments and 6 individual homes, all in the city of Funchal. This deployment lasted until the end of November when a revised version of the feedback was installed in the same 17 apartments and 6 individual homes. This second deployment lasted until April of 2012 when the system was removed from the participant homes. In the meantime, during this period three apartments and one house had their system removed earlier either for technical difficulties (one house and one apartment) or simply because the householders no longer wanted to participate in the study (two apartments).

The first and second deployments were done using a preliminary version of our sensing hardware and software platform for energy monitoring and feedback. These consisted of one sensor for current, another for voltage and a netbook that served both for measuring the energy usage and to provide feedback to the householders. The netbook and the sensors were installed in the main power feed (Figure 1 – left), thus covering the entire house consumption. Current and voltage were continuously sampled at 8000 Hz using the netbook built-in soundcard, and the consumption metrics (apparent, real and reactive power) calculated using the equations 1, 2 and 3 respectively, where S is apparent power, P is real power, Q is reactive power, I_{RMS} is the average current, V_{RMS} is the average voltage and ϕ is the phase angle between the instantaneous current and voltage measurements.

$$S = I_{RMS} * V_{RMS} \quad (1)$$

$$P = S * \cos(\phi) \quad (2)$$

$$Q = S * \sin(\phi) \quad (3)$$

These metrics were calculated at a rate of 50 samples per second (the mains frequency of the monitored houses is 50 Hz), and subsequently used for event detection which is the process of identifying the changes in the total load that happen in response to electric devices changing their working mode (e.g. a device turning *on* or *off*). In the meantime all the metrics were

stored in a local database (aggregated at 1 measurement per minute) along with the detected power changes for feedback and future data analysis purposes.

The eco-feedback was given using the built-in display of the netbook (Figure 1 – right) through an external application that provided historical data loaded from the local database and real time information on energy consumption by directly connecting to the energy monitoring software. To lower the cost of the deployments the data was acquired via the microphone jack and the audio board acting as the Data Acquisition (DAQ) components of the system. Additionally, the feedback software kept a log of every interaction between the householders and the eco-feedback by keeping track of the mouse clicks on the user interface.

Fig. 1. Energy monitoring and feedback platforms – version 1. The sensors were installed in the main power feed (left) along with the netbook that was used to perform all the calculations and provide feedback to the residents (right).

The third deployment lasted from August 2012 to January 2013. During this period we monitored the consumption of 17 apartments in three building blocks of a recently build condominium in the city of Funchal. Lastly, a fourth ongoing deployment started in July 2013 with 10 apartments from a single building close to our campus.

Our hardware setup evolved according to the limitations of the different deployments. Our initial setup was located in the mains and raised some issues of limited accessibility by some household members (especially kids) and some questions about the security of the system. We consequently made some changes in the original monitoring framework. The most significant change involved replacing the netbook with a more capable DAQ system that was now installed in the main lobby of the apartment buildings where all the meters from the electric company are mounted (Figure 2 – left). This enabled us to measure the energy consumption of multiple homes from a single sensing location. All the required signals were acquired and processed by a single computer that was also responsible for providing access to the data through a web services layer.

Regarding the eco-feedback, all the visualizations were implemented in a 7'' tablet (Figure 2 – right) that gathered data using either the available web services or a custom communication protocol that was preserved across versions. Additionally, the new feedback software also kept track of the entire user interactions by keeping local logs of all the transitions between the different screens.

Fig. 2. Energy monitoring and feedback platforms – version 2. All the sensors were installed in the lobby of the building (left), and the feedback was provided using a 7'' tablet (right).

III. SustData Dataset

The SustData dataset contains all the data collected during the four deployments described in the previous section, namely energy consumption, power events and user events, to which we also added the demographics of all the monitored houses. Additionally, we also created a record of the electric energy production by the local electric company and compiled the weather information during all the deployments from a weather web service.

A. Energy Consumption

The energy consumption data is a record of several energy consumption related metrics, e.g. real, reactive and apparent power aggregated at one minute time intervals, from the 50 homes in our deployment (44 apartments and 6 individual houses). Table I lists all the energy consumption measurements that are available in SustData.

TABLE I. Energy Consumption Measurements.

Field	Description	Units
home_id	Monitored home unique identifier	-
timestamp	Date and time of the measurement	datetime
deploy	Deployment identifier	-
Imin	Minimum current	A
Imax	Maximum current	A
Iavg	Average current	A
Vmin	Minimum voltage	V
Vmax	Maximum voltage	V
Vavg	Average voltage	V
Pmin	Minimum real power	W
Pmax	Maximum real power	W
Pavg	Average real power	W
Pavg_t	Average Real Energy	Wh
Qmin	Minimum reactive power	VAR
Qmax	Maximum reactive power	VAR

Field	Description	Units
Qavg	Average reactive power	VAR
Qavg_t	Average reactive energy	VARh
Smin	Minimum apparent power	VA
Smax	Maximum apparent power	VA
Savg	Average apparent power	VA
Savg_t	Average apparent energy	VAh
PFmin	Minimum power factor	-
PFmax	Maximum power factor	-
PFavg	Average power factor	-

B. Power Events

The power events are extracted from the raw power measurements (active power at 50 Hz) by applying a modified version of the change of mean detector algorithm presented in [15], where the authors apply a statistical method based on the Generalized Likelihood Ratio test (GLR) to find the likelihood of a potential change of mean occur between two consecutive time periods.

Our change of mean detector works with one sliding window (*detection window*) that is used to calculate the likelihood of a change of mean happen at a given sample. A second sliding window, called *voting window*, is used to find the *extrema* values of the likelihood test. The detection window [*i, k*] is composed by two separate windows, [*i, j*[and [*j, k*], *pre-event* and *post-event* respectively. For each sample in the power signal the likelihood of a power change occurring at that instant is given by equation 4:

$$l(x) = \frac{\mu_{[i,\ j[} - \mu_{[j,\ k]}}{\sigma^2_{[i,k]}} \times \left| P(x) - \frac{\mu_{[i,j[} + \mu_{[j,k]}}{2} \right| \qquad (4)$$

Where $\boldsymbol{\mu_{[i,\ j[}}$ and $\boldsymbol{\mu_{[j,\ k]}}$ are the mean of the pre-event and post-event windows respectively, $\boldsymbol{\sigma_{[i,\ k]}}$ is the standard deviation of the detection window and $\boldsymbol{P(x)}$ *is* the active power of the $\boldsymbol{x^{th}}$ sample.

Figure 3 illustrates the power event detection process, and it is possible to see that when a power change occurs (bottom line in blue), this will be reflected in the maximization (positive power changes) or minimization (negative changes) of the events likelihood (top line in red). Therefore, the voting window of our algorithm works by simply finding the *maxima* and *minima* (i.e. the *extrema*) of the event likelihood results that correspond to the positions of the power events in the aggregated power signal.

This algorithm has four parameters: the sizes of the pre-event, post-event and voting windows and the minimum step change of interest bellow which the likelihood of an event occurring is set to zero.

Before deploying the system we performed laboratory tests [16] to find the values for this parameters that would give the best results in terms of detected power changes. In the end the best results (with a percentage of True Positives above 95%) were obtained with the following values: pre-event window size: 150, post-event window size: 100, voting window size: 50 (all in samples) and the minimum step change of interest was set to 30 Watts, that were then used in the four deployments. Furthermore posterior evaluation of the algorithm using this values against one of the datasets previously mentioned (BLUED) showed an average sensitivity (ability to detect positive results) around 90%, which was in accordance with our initial test results.

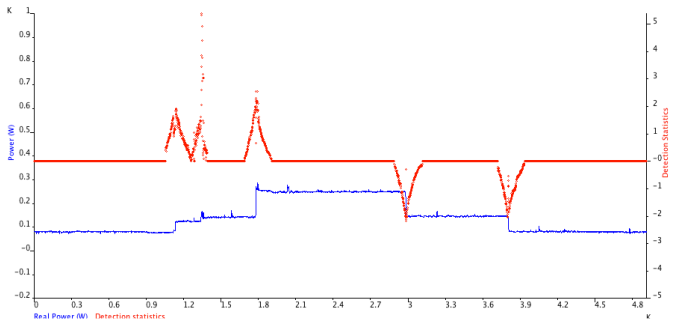

Fig. 3. Event detection process. As the changes in active power (bottom line) happen they will be reflected in the maximization or minimization of the event likelihood statistics (top line).

The power event data is a record of all the power changes with a mean real power change of at least ± 30 Watts. Table II lists the measurements that characterize each power event.

TABLE II. Power event measurements

Field	Description	Units
home_id	Unique identifier of the monitored home	-
timestamp	Date and time of the measurement	datetime
deploy	Deployment identifier	-
delta_P	Real power change	W
delta_Q	Reactive power change	VA
trace_P	Real power trace of the event (50 Hz)	W
trace_Q	Reactive power trace of the event (50 Hz)	VAR

The event trace is the collection of all the power values in the vicinity of the power event, as shown in figure. 4. In our case the power event traces contain 150 measurements before the change and 100 after, which at 50 Hz correspond to 3 and 2 seconds respectively.

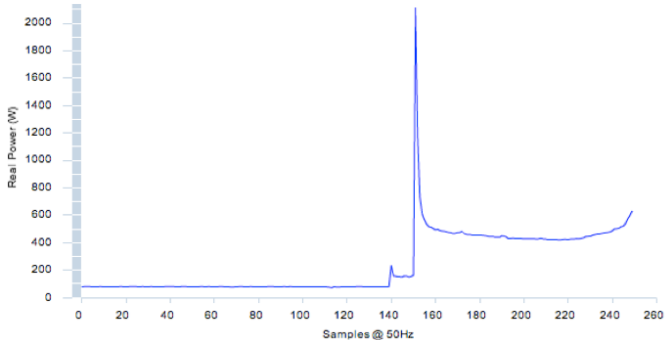

Fig. 4. Real power transient of a microwave turning *on*.

C. Interactions with the Eco-feedback

This is a record of the frequency at which the householders used the feedback. The interactions with the system were only collected for houses with feedback devices (some of the houses were used as control group with no feedback). Also no interactions are available during the baseline periods of the different studies (period where electricity consumption was collected but no feedback was provided). Table III lists the features that describe each user event.

D. Demographics

The demographics data is a record that describes the participating families, their homes and the periods for which consumption and user event data is available. Table IV lists the demographics features of our dataset.

E. Environmental Data

Given the importance of weather phenomena of the production of energy, we have also collected extended information on this. The data was collected from an online repository[2] of environmental data, and consists of several measurements (listed in table VI) collected at 30 minute read intervals.

F. Electricity Production

Since the 18th of June 2013 we are also recoding the numbers of the electricity production using a web service made available by the local electric company. The production data is recorded at fifteen minute read intervals and besides the overall value, the disaggregated production by source is also recorded. Table V lists the production measurements that are available.

G. Additional data

In addition to the data that we have already described we are also making available some complementary items that can be relevant for future research: i) individual householder demographics, including gender, age, occupation, among others; ii) list of appliances in some of the participating houses, and iii) a time dimension table, supplemented with daily weather information, e.g. minimum and maximum temperatures, sunrise and sunset times.

TABLE III. User event features

Field	Description
home_id	Unique identifier of the monitored home
timestamp	Date and time of the measurement
deploy	Deployment identifier
type	Type of interaction. Either mouse or touch
view_id	Identifier of the visualized screen
view_name	Name of visualized screen

TABLE IV. Demographic features

Field	Description	Units
home_id	Monitored home unique identifier	-
building_id	Building identifier	-
begin_monitoring	Date and time of the first measurement	datetime
end_monitoring	Data and time of the last measurement	datetime
begin_feedback	Date and time of feedback deployment	datetime
end_feedback	Date and time of feedback removal	datetime
type	Type of residence. Apartment or house	-
bedrooms	Number of bedrooms	-
adults	Number of adults	-
children	Number of children	-
contacted_power	Contracted power with the provider	kWh

TABLE V. Electric energy production measurements

Field	Description	Units
timestamp	Date and time of the measurement	datetime
total	Total production	MWh
thermal_fuel	Electricity produced by burning fuel	MWh
hydro	Hydro electricity produced	MWh
eolic	Wind farms production	MWh
photovoltaic	Solar electricity produced	MWh
thermal_waste	Electricity produced by burning waste	MWh

TABLE VI. Environmental data measurements

Field	Description	Units
timestamp	Date and time of the measurement	datetime
temperature	Outside temperature	ºC
humidity	Relative humidity	%
pressure	Relative pressure	hPa
wind_dir	Wind direction	-
wind_speed	Wind speed	km/h
precipitation	Precipitation levels	mm
events	Relevant events, e.g. rain or thunder	-
conditions	Sky conditions. e.g. partly cloudy	-

[2] www.wunderground.com

IV. SustData Characterization

The SustData dataset contains, at the time of this writing (March 10, 2014), over 50 million individual records of electric energy related data, spanning a total of 1144 distinct days since the 29th of July 2010.

A. Energy consumption

As mentioned above, our dataset contains over 50 million individual records, from which almost 25 million are individual power readings as summarized in table VIII.

TABLE VII. Energy consumption data summary

Dep	Samples	Days	Min. Days	Max. Days	Min Date (YY/MM/DD)	Max Date (YY/MM/DD)
1	3.474.557	123	51	119	2010/07/10	2010/11/10
2	12.481.536	504	240	511	2010/11/25	2012/04/20
3	5.671.576	298	237	297	2012/08/01	2013/05/25
4	2.884.512	219	187	217	2013/07/31	2014/03/10
	24.512.181	**1144**	---	**1144**	---	---

The first deployment was the shortest one with 123 different days, while the second is the longest with consumption data for 504 days. The minimum number of days (**Min. Days**) is, as the name suggests, the smallest number of days available for a single house (reciprocate for the maximum number of days).

Each individual day contains 1440 measurements for every monitored house. Figure 5 shows the graphical representation of the real and reactive power from a random day in one of the dataset houses, in which one can easily notice peaks of consumption in very specific time periods, namely, early morning, noon and late afternoon / early evening.

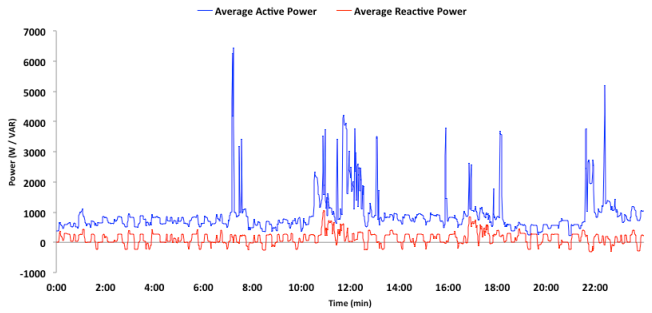

Fig. 5. Real and Reactive power traces for one full day of consumption in one random house of the SustData dataset. *(Best viewed in color)*

B. Power Events

Currently the dataset contains over 11 million individual power events across all the four deployments, as summarized in table VIII:

TABLE VIII. Power events summary

Dep	Events	Min. Events	Max. Events	AVG	SD	Day AVG	Day SD
1	1.487.945	17.656	203.336	70.855	48.834	897	625
2	6.057.701	51.564	722.813	288.462	171.532	685	380
3	1.822.903	18.610	495.596	130.207	153.872	460	525
4	1.745.232	58.315	485.933	174.523	122.640	880	610
	11.113.781	---	---	---	---	---	---

Where **min. events** is the minimum number of power events in a single house (reciprocate to max. events), **AVG** is the average number of power events between all the houses, **Day AVG** is the daily average of power events and **SD** is the standard deviation (**Day SD** is the daily standard deviation).

A quick inspection of the results immediately reveals the high values for the daily standard deviation, which is a clear indicator of the large difference in the number of power events across the different houses, for instance, in the second deployment four houses have over 1000 daily power events in average, while seven houses have less than 300 power events per day.

C. Interactions with the Eco-feedback

Regarding the interactions with the eco-feedback our dataset contains 14027 individual records from a total of 32 households (in the third and fourth deployments some houses were part of the control group, 11 and 5 respectively, and no eco-feedback was provided).

Table IX shows a summary of the eco-feedback interactions data. Once again the values for the standard deviation are high (with the exception of the current deployment), which is an indicator that not all the household members used their systems with the same frequency. In fact, some of the houses have very low numbers of interactions in the first three deployments (3, 14 and 20).

TABLE IX. Summary of the interactions with eco-fedback data

Dep	Houses	Interactions w/ feedback	Min. User Events	Max. User Events	AVG	SD
1	21	3.739	3	548	178	155
2	21	7.682	14	928	366	226
3	6	1.424	20	732	237	247
4	5	1.182	136	356	236	72
	---	**14.027**	---	---	---	---

D. Environmental and Electricity Production Data

The SustData dataset contains, at the time of this writing, 24886 records of electricity production data, recorded at 15-minute intervals since the 18th of June 2013 (264 days so far). Additionally we have also collected individual environmental data measurements for each of the 1144 days in our dataset at a rate of 48 measurements per day (54912 records).

V. Dataset Implementation

The SustData dataset is a free and publicly available dataset for all researchers to use and can be accessed from *http://aveiro.m-iti.org/data/sustdata*. Our dataset is persisted using a Not only SQL (NoSQL) Database Management System (DBMS), and the available data can be accessed using its open Application Programming Interface (API).

A. Data Persistence

The most common way of persisting publicly accessible data continues to be, after so many years, using text files that follow a certain structure that is then given to the users (normally in the form of readme files). But, while this seems to be the best option for the dataset creators it presents limitations for the actual datasets users, for instance, there is no way of having a preview on the data without downloading the files completely.

With this in mind, and given the size and variety of the data in our dataset, we have opted to persist it using a NoSQL DBMS, that are known for being capable of dealing well with large amounts of data in opposition to traditional relational DBMS that tend to suffer serious performance issues as the database grows in size. Moreover, NoSQL DBMS are optimized for retrieve and append operations, which are the operations that we are particularly interested in.

In our case we selected the MongoDB[3] implementation of NoSQL, which besides the aforementioned advantages also offers a query language to create "SQL-like" queries that are an important feature of our API.

B. Web Server and Web API

The server-side of our system was implemented using the Node.js[4] software platform to setup and run a web-server, Mongoose[5] library for data modeling and query building, and the Express[6] web application framework to create our web API and the dataset explorer web application. The combination of MongoDB, Node.js, Mongoose and Espressjs was selected due to their seamless integration that results from the fact that they all share the use of JavaScript (MongoDB, Node.js and Expressjs are all implemented using this programing language) and the JavaScript Object Notation (JSON[7]) that is the standard data format used by MongoDB.

Regarding the API, it offers a set of pre-defined functions to access and manipulate the several data types in our dataset, including aggregation, grouping, sorting and filtering. These features were designed and implemented to allow quick exploration of the existing data as well as to provide mechanisms to quickly export the data to common text formants, thus proving an easy way to create subsets of the original dataset for offline operation, instead of forcing users to download the whole dataset.

Finally, interested users can access the datasets using one of three options available: 1) *online* using the dataset explorer web application, 2) *online* by directly accessing the web API HTTP methods for retrieving data and 3) *offline* after exporting the data using the export features available in the dataset explorer.

[3] www.mongodb.org

[4] www.nodejs.org

[5] www.mongoosejs.com

[6] www.expressjs.com

[7] www.json.org

VI. Potential Uses

Sustainable energy generation, distribution and consumption can greatly benefit from the potential offered by ICT. In this section we describe some possible uses of our dataset towards enabling more intelligent energy management systems.

A. Eco-feedback Research

A key aspect of energy monitoring effectiveness is how the information is presented to the users and there is a whole field of research devoted to this. It is known as eco-feedback technology [17], and focuses on finding what kinds of residential energy feedback are most effective and appropriate in specific contexts and locations.

Information visualization is the core of any eco-feedback approach and we believe that this dataset can play an important role in creating better eco-feedback experiences. For example, the consumption data can be used as inputs for different interface designs to help answer critical research questions in this field related to temporal resolution (minute, hour, day, week, month and year) and presentation mode (real-time and historical). Moreover, since we have data for several houses, each one with different consumption patterns, it is possible to understand how the feedback systems will behave in different situations, e.g. high consumption vs. low consumption families.

The same data can also be used to evaluate more abstract visualizations that are harder to test or predict how they will behave in different consumption scenarios. This is the case of most ambient and artistic visualizations where researchers attempt to integrate consumption information in the domestic environment [18]. Furthermore, having long periods of consecutive consumption data can be used to find energy usage patterns that can be used to create novel eco-feedback experiences like the one in [19] where the authors attempted to leverage the emotional connection between energy consumption and the local endemic landscape by combining real-time and historical baseline consumption to change the landscape according to the observed consumptions patterns.

Similarly the recorded interactions with the feedback can be useful to understand how this was used over time, assess which features the users preferred and find navigation patterns that can guide designers in the early stages of designing eco-feedback interfaces.

Finally the power event data can be used to create and evaluate eco-feedback systems designed for labeling power events with the respective appliance names to be used as training data for load disaggregation algorithms. Having humans manually labeling sensor data is a problem transversal

to every supervised machine learning approach, and with our dataset it is possible to combine consumption and power events to design and test interfaces that can tackle both the problem of motivating the consumers to perform this task, while increasing the accuracy of the provided labels.

B. Prediction and Forecasting

Improved power plant efficiency was largely identified as an area where significant reductions in CO_2 emissions can be achieved using only already established techniques [20]. One of these techniques is energy demand management that refers to the ability to balance the supply of electricity on the network by adjusting the load in order to reduce peak demand periods.

The ability to accurately predict future energy needs is cornerstone in proper demand side management, and many research efforts have been devoted to this in the last couple of years [21] including contributions from different fields like time series and regression analysis, artificial neural networks, genetic algorithms and fuzzy logic.

Some of the investigated methods rely heavily on past consumption data to predict future demand, and therefore we argue that our dataset can be of added value in this situations, especially given the high granularity of our data (one measurement per minute) that can be easily manipulated to test different prediction periods (e.g. hour, day, week), and the considerable number of different houses (50), that can be used to evaluate the outcomes of the prediction algorithms in a variety of energy consumption scenarios.

Furthermore, the power events data can also be important to investigate forecasting approaches that look at individual device uses to create prediction models of the overall consumption. These approaches, commonly denominated bottom-up approaches, are based on the obvious relationship between the overall consumption and the consumption patterns of each individual appliance. Figure 6 shows this relationship in our dataset with highlights in the areas where it is possible to see the direct relationship between the actual power and the load changes.

Not less important than energy demand management is the smart grid initiative that is significantly increasing the fraction of grid energy that is contributed by renewables. Nevertheless, the high volatility and unpredictable nature of renewables increases the difficulty of integrating them into the grid without affecting its stability. Thus, the ability to accurately predict future renewable generation is important since the grid must be able to dispatch mechanisms to compensate over or under generation from renewable sources.

It is therefore not surprising that researchers are also devoting their efforts into creating forecasting models for energy generation from renewable sources based on environmental data such as solar penetration and wind speed [22], [23]. Such prediction models are created based on historical production and environmental data records, thus making our dataset suitable for creating and evaluating these models using the provided renewables and temperature records, as shown in figures 7 and 8 where we plot one day of environmental information against production data.

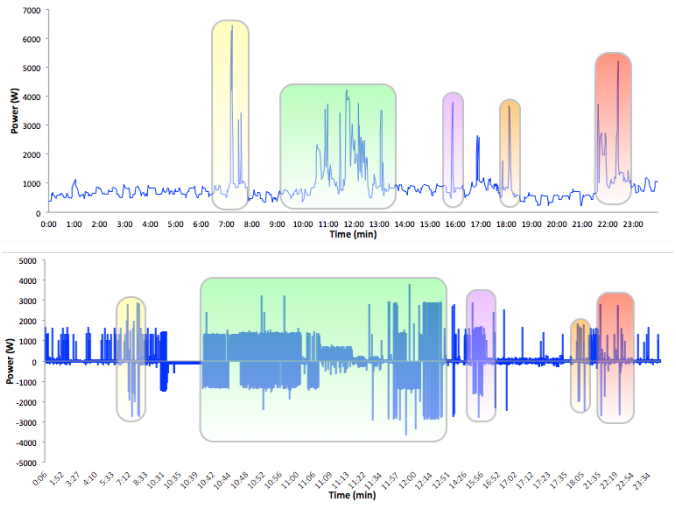

Fig. 6. Two representations of the power consumption: on the top using the real power one-minute interval measurements, and on the bottom using the real power change of the events that were detected during that day. *(Best viewed in color)*

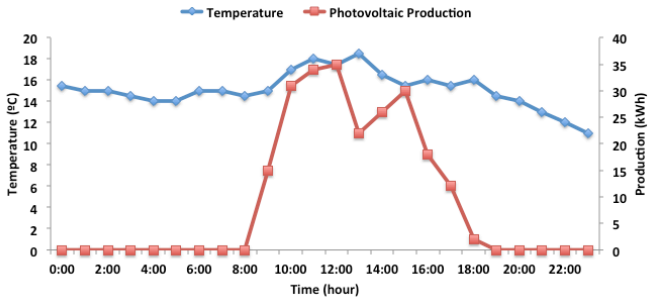

Fig. 7. Environmental and production data comparison: temperature against photovoltaic production of electricity.

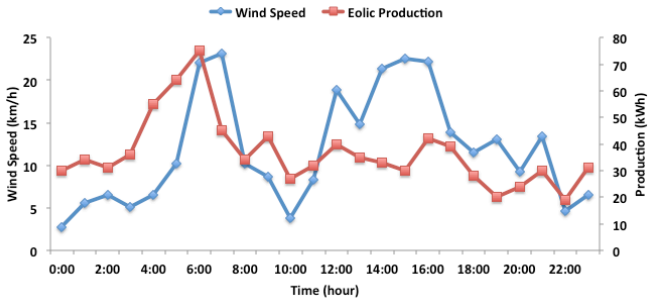

Fig. 8. Environmental and production data comparison: temperature against eolic production of electricity.

C. Unsupervised Feature Learning for Load Disaggregation

Inspired by recent discoveries in the field of neuroscience a new subfield of machine learning as emerged. This is the subfield of deep learning [24] and, in very high-level terms, the main goal of deep learning architectures is to learn the best possible feature representations of the examples in the data.

One of the most promising characteristics of deep learning architectures is the possibility of learning features from unlabeled data given that the algorithms are provided with large quantities of unlabeled examples from which they can

learn good feature representations. This is known as unsupervised feature learning and it is in this area that our dataset can play a very important role in load disaggregation since we are now releasing large amounts of data for both continuous power consumption at 1-minute intervals and power events with real and reactive transients at 50 Hz that can be used to extract relevant features that can ultimately improve the overall energy disaggregation results.

VII. Conclusion

In this paper we have introduced SustData, a public dataset for sustainable electric energy research. Ultimately, our goal with the release of this dataset is to provide the research community with a complete set of data that can be used in the development and evaluation of novel techniques to improve the already exiting energy monitoring and forecasting solutions.

As it was previously mentioned, this dataset is still being updated with data from our most recent energy monitoring deployment, and future work includes adding data from other sources, including the addition of labeled power event data, thus making this dataset suitable for load disaggregation research. Furthermore, we are also very excited with the recent prospects of deep-learning architectures offering the possibility of extracting features from unlabeled data, which we believe can represent enormous advances in this very active field of research.

In summary, we are convinced that having more and better publicly available data can further motivate the sustainability research community to apply state of the art machine learning and data mining techniques to tackle the problem of excessive carbon emissions due to unsustainable electricity consumption.

Acknowledgements

This research was partially funded by the CMU | Portugal SINAIS project (CMU-PT/HuMach/0004/2008) and the Portuguese Foundation for Science and Technology (FCT) doctoral grant SFRH/DB/77856/2011. Finally, we would like to acknowledge all the participating families and the local electric company of Madeira (EEM) for their technical support.

References

[1] International Energy Agency, "2013 Key World Energy Statistics," International Energy Agency, 2014.

[2] U.S. Energy Information Administration, "International Energy Outlook 2013," U.S. Energy Information Administration, Jul. 2013.

[3] L. J. Becker, "Joint effect of feedback and goal setting on performance: A field study of residential energy conservation," *J. Appl. Psychol.*, vol. 63, no. 4, pp. 428–433, 1978.

[4] C. Fischer, "Feedback on household electricity consumption: a tool for saving energy?," *Energy Effic.*, vol. 1, no. 1, pp. 79–104, Feb. 2008.

[5] L. Pereira, F. Quintal, N. Nunes, and M. Bergés, "The Design of a Hardware-software Platform for Long-term Energy Eco-feedback Research," in *Proceedings of the 4th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, Copenhagen, Denmark, 2012, pp. 221–230.

[6] N. Batra, M. Gulati, A. Singh, and M. B. Srivastava, "It's Different: Insights into Home Energy Consumption in India," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, New York, NY, USA, 2013, pp. 3:1–3:8.

[7] D.J. Cook, M. Schmitter-Edgecombe, Aaron Crandall, Chad Sanders, and Brian Thomas, "CASAS Project." .

[8] K. Bache and M. Lichman, "Individual Household electric power consumption dataset." Irvine, CA: University of California, School of Information and Computer Science., 2013.

[9] Sean Barker, Aditya Mishra, David Irwin, Emmanuel Cecchet, and Prashant Shenoy, "Smart*: An Open Data Set and Tools for Enabling Research in Sustainable Homes." .

[10] Z. Kolter and J. Matthew, "REDD: A public data set for energy disaggregation research," presented at the Workshop on Data Mining Applications in Sustainability (SustKDD), San Diego, CA, USA, 2011.

[11] K. Anderson, A. Ocneanu, D. Benitez, D. Carlson, A. Rowe, and M. Berges, "BLUED: A Fully Labeled Public Dataset for Event-Based Non-Intrusive Load Monitoring Research," presented at the Workshop on Data Mining Applications in Sustainability (SustKDD), Beijing, China, 2012.

[12] S. Makoni, F. Popowish, L. Bartram, B. Gill, and I. V. Bajic, "AMPds: A Public Dataset for Load Disaggregation and Eco-Feedback Research," *Electr. Power Energy Conf. EPEC*, pp. 1–6, 2013.

[13] C. Holcomb, "Pecan Street Inc.: A Test-bed for NILM," presented at the 1st International NILM Workshop, Pittsburgh, PA.

[14] J. Kelly and W. Knottenbelt, "`UK-DALE': A dataset recording UK Domestic Appliance-Level Electricity demand and whole-house demand," *ArXiv14040284 Cs*, Apr. 2014.

[15] D. Luo, L. K. Norford, S. B. Leeb, and S. R. Shaw, "Monitoring HVAC Equipment Electrical Loads from a Centralized Location Methods and Field Test Results," *ASHRAE Trans.*, vol. 108, pp. 841 – 857, 2002.

[16] L. Pereira and N. J. Nunes, "Low cost framework for non-intrusive home energy monitoring and research," in *Proceedings of the 1st International Conference on Smart Grids and Green IT Systems*, Porto, Portugal, 2012, vol. 1.

[17] J. Froehlich, L. Findlater, and J. Landay, "The Design of Eco-feedback Technology," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Atlanta, GA, USA, 2010, pp. 1999–2008.

[18] J. Rodgers and L. Bartram, "Exploring Ambient and Artistic Visualization for Residential Energy Use Feedback," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2489–2497, Dec. 2011.

[19] F. Quintal, L. Pereira, N. Nunes, V. Nisi, and M. Barreto, "WATTSBurning: Design and Evaluation of an Innovative Eco-Feedback System," in *Human-Computer Interaction – INTERACT 2013*, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds. Springer Berlin Heidelberg, 2013, pp. 453–470.

[20] Pacala and R. Socolow, "Stabilization wedges: solving the climate problem for the next 50 years with current technologies.," *Science*, vol. 305, no. 5686, pp. 968–972, 2004.

[21] L. Suganthi and A. A. Samuel, "Energy models for demand forecasting—A review," *Renew. Sustain. Energy Rev.*, vol. 16, no. 2, pp. 1223–1240, Feb. 2012.

[22] N. Sharma, P. Sharma, D. Irwin, and P. Shenoy, "Predicting solar generation from weather forecasts using machine learning," in *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2011, pp. 528–533.

[23] A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh, "Current methods and advances in forecasting of wind power generation," *Renew. Energy*, vol. 37, no. 1, pp. 1–8, Jan. 2012.

[24] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]," *IEEE Comput. Intell. Mag.*, vol. 5, no. 4, pp. 13–18, Nov. 2010.