GreenDataNet

D2.5 – Forecasting Algorithm for Energy Consumption

Status
[Final]
Authors
Prof. Brunelli Davide – UNITN
Prof. Petri Dario, Maurizio Rossi, Ivan Minakov and Davide Sartori – UNITN
Rev 0.5
Contributors:
-
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLE OF CONTENTS</td>
<td>2</td>
</tr>
<tr>
<td>REVISION SHEET</td>
<td>3</td>
</tr>
<tr>
<td>KEY REFERENCES AND SUPPORTING DOCUMENTATIONS</td>
<td>4</td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>6</td>
</tr>
<tr>
<td>1.1 Document Purpose</td>
<td>7</td>
</tr>
<tr>
<td>1.2 Document overview</td>
<td>8</td>
</tr>
<tr>
<td>2. RELATED WORKS</td>
<td>9</td>
</tr>
<tr>
<td>3. ENERGY FORECASTING</td>
<td>11</td>
</tr>
<tr>
<td>3.1 Forecasting approaches</td>
<td>11</td>
</tr>
<tr>
<td>3.2 Data-sets</td>
<td>13</td>
</tr>
<tr>
<td>4. RESULTS</td>
<td>17</td>
</tr>
<tr>
<td>5. PORTABILITY</td>
<td>22</td>
</tr>
<tr>
<td>6. CONCLUSIONS</td>
<td>23</td>
</tr>
<tr>
<td>Revision Number</td>
<td>Date</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------</td>
</tr>
<tr>
<td>Rev 0.1</td>
<td>28/10/2015</td>
</tr>
<tr>
<td>Rev 0.4</td>
<td>30/11/2015</td>
</tr>
<tr>
<td>Rev 0.5</td>
<td>9/12/2015</td>
</tr>
</tbody>
</table>
KEY REFERENCES AND SUPPORTING DOCUMENTATIONS


[Online] https://nwsc.ucar.edu/green


**INTRODUCTION**

Data centers have been rapidly increased in number during the last decade and this trend will continue in the next future, because of the rise of cloud computing, the huge growth of mobile devices, the continuing development of new IT services and IoT applications [1]. Along with market trends, environmental awareness and energy savings are becoming a very relevant issue for data centers managers and regulatory bodies, since, 10% of the global consumption of electrical energy has been estimated to be consumed by IT infrastructures [2]. To tackle the electrical energy demand of the IT infrastructures many solution are currently under investigation both in academia and industry, like new computing hardware based on low power ARM architectures [3] and “green” data centers that exploits renewable energy sources to mitigate the load on the electricity Grid [4]. The most valuable solutions are envisioned in the design of mega data centers of the future [5].

So far, all the bigger companies in the IT market (Amazon, Google, Apple...) have already introduced renewable energy sources in the supply chain of their infrastructures in different ways, for example wind and solar power plant are used to generate green energy while free cooling helps to reduce (sometimes even replace) the utilization of conditioning systems (HVAC) at those latitudes where outside temperature is below 20° Celsius all the year.

Energy consumption forecasts are required to optimize the use of renewable energy resources, through auxiliary energy storage systems, not only in data centers, and many approaches has been proposed in literature to solve renewable integration in such management systems [6].

Renewable energy sources are unpredictable by nature but in many cases, estimating their short-term trend (one day ahead) with small error (Mean Average Percentage Error – MAPE – close to 10%) is possible, as it has been demonstrated in [7] and also discussed in D3.4. Similar results can be expected when dealing with electricity demand prediction at building scale (few tens of KWs) [8].

While forecasting methods of renewable energy sources can be used for a broad range of applications (e.g. domestic and industrial plants, standalone embedded monitoring systems, etc.); prediction techniques of the energy consumption are more related to specific target requirements.

So far, Datacenter’s environment has gathered few efforts for the prediction and the analysis of the electricity consumption because most of the time IT infrastructures run at constant performance level to guarantee uninterrupted availability of hosted services.
Electricity consumption in this case results almost flat, at a first glance, and at constant level every day of the year. For this reason much simpler approaches have been used so far. This characteristic of data center is going to change in the next future with the large scale introduction of low-power/high-performance computing units with highly varying energy requirements, for example servers that can be dynamically switched off according to the workload (more details will be provided in the following) and this will determine unexpected peaks or shift levels. In this case the importance of tailored forecasting algorithms will be the key to achieve optimal energy management.

Predictions of electrical energy demand are gaining a lot of attention also from electricity Grid managers. The transition to the smart grid (namely the transformation of the electrical energy transmission and distribution grids into a self-monitoring/-healing network where distributed generation, bi-directional flows of energy and supply-demand balance will be automatically adjusted) requires a set of new instruments to enable the unsupervised, machine based management, starting from a more precise state estimation. Along with new measurement instrumentation (e.g. Phasor Measurement Units) and control algorithms the state estimation can significantly take advantage from so-called pseudo-measurements like predictions and historical data [9].

1.1 DOCUMENT PURPOSE

This deliverable describes how to tackle the problem of forecasting energy consumption of data centers considering stochastic methods for time-series analysis. Starting from the specification provided by work package 1 (WP1), in particular on the software architecture specification (D1.5) and the results obtained in the same work package (WP2), we evaluated algorithms that can be integrated in the software environment of GDN (as a module) that exploit the historical data collected by the monitoring system (D2.1) and will provide useful information that will be used for Smart Grid integration (D3.13) along with data coming from other modules (for example PV production forecasting as described in D3.4).

We selected the Holt-Winters (HWT) forecasting algorithm and compared its performance with the traditional persistence approach currently used in literature. This exponential smoothing technique was demonstrated as optimal solution for a very general class of state-space models, broader than the traditional autoregressive methods, which have higher memory requirements too. Additionally, we evaluated an extended HWT implementation that was demonstrated in literature [8]. We considered power traces of real data centers provided by GreenDataNet project’s partners, which are characterized by different dimensions and exhibit different features. The results achieved by the pro-
posed algorithm show better performance than the reference method with highly variable workloads, where the latter one faces significant limitations. Comparable results have been obtained in all the other cases.

The algorithm, opportunely shaped, will be then implemented in the final demonstrator as a software module to facilitate the prediction of the power/energy consumption of different data center subsystems as highlighted in the project’s outline depicted in Figure 1.

![Network of Data Centres / Smart Grid](image)

**Figure 1. The forecasting algorithm role in the GDN project framework**

### 1.2 DOCUMENT OVERVIEW

The document is organized as follows, Sec. 2 briefly introduces the state of the art in the related fields, Sec. 3 describes the data-sets used and the forecasting algorithms under comparison while Sec. 4 presents the simulations results’. Finally Section 5 illustrates the interface of the SW library for integration in other GDN tools, while Section 6 concludes the deliverable.
2. RELATED WORKS

Despite the huge number of attempts presented in the last five years about modeling local and distributed data center resources, not enough efforts have been dedicated to analyze forecasting methods for energy consumption prediction. Researchers in this field target their analysis to the prediction of CPU utilization and workload evolution in time, to optimize the usage of resources employing Dynamic Voltage and Frequency Scaling (DVFS) techniques. In this case a prominent approach is to use Kalman filters [11] which has been demonstrated to be effective in several works [12].

CPU utilization however, is not directly related with energy consumption of the data center and it is not the most power hungry component in the system since cooling, power transformation, distribution and management subsystem and the network infrastructure are examples of continuously running equipment with high energy demand. To measure the energy performance of a data center the “Green Grid”, an association of IT organizations focused on making data centers more energy efficient, introduced the Power Usage Effectiveness index which is today the de-facto standard performance assessment indicator for data centers. Even if the PUE of future data center generations is targeted to be slightly higher than one (1.0x) (thanks to low power electronics, free cooling and resource optimization), today most of the data centers around the world run in the range 1.5 to 2.0 PUE [13].

To achieve these goals and reduce today the PUE to around 1.3 as the GreenDataNet target, effective energy resources management must be implemented and forecasting is an essential tool as described in [4], [14]. The articles present a pilot “green” data center which is equipped with the last generation ARM architecture servers. The pilot has been equipped with a small photovoltaic (PV) plant and a combined HVAC/free-cooling system to demonstrate the advantages of such integration of new technologies and to test their state-of-art energy management algorithm. The controller optimizes the renewable energy usage exploiting an energy buffer (an improved UPS system made with Li-ion batteries) and shaping the execution speed of the workload. The optimization algorithm implements different forecasting tools for workload, renewable energy and total energy consumption of the data center. However the energy consumption forecasting module is based on the simplistic “persistence” approach, where it is assumed that the values expected for the next time-horizon is exactly the same of the previous one.

As far as we know this is the most widely used approach for this task in data center scenario, since energy consumption exhibits stationary features.
We will show that this is not always true, data centers energy consumption may vary significantly depending on the utilization (cloud services against high-performance computing), thus dedicated and specialized processing tools are necessary to achieve optimal performance in the prediction.
3. ENERGY FORECASTING

In this section we summarize the two forecasting algorithms evaluated in the comparison, then we briefly discuss the four data-set used, presenting their features and specific characteristics, before entering into the details of the simulation results’.

3.1 FORECASTING APPROACHES

As introduced in Sec. 3, the most used method to predict energy consumption in data center management application is the so called “persistence” approach, where past samples are simply replicated to generate the forecast. This approach has great performance for stable and constant consumption profiles. However the performance rapidly decreases if in the time-series exist some periodic patterns that are different from the forecasting horizon. For example, corporate buildings have high energy requirements during working days and almost none during weekends, if the forecast is computed on a daily basis we can expect large errors during both transitions from weekdays to weekends and vice versa. The same applies to data centers, for example companies may decide to run specific tasks (market analysis, consolidation of the data, etc.) on a fixed working day or monthly, generating a specific pattern that cannot be predicted with persistence.

This scheme has been compared with an algorithm from the class of exponentially weighted moving average (EWMA) methods, namely the Holt-Winter exponential smoothing technique (HWT in the following). This algorithm was developed more than sixty years ago to analyze the stock exchange market. Since then, many researchers have worked with this scheme and nowadays several versions exists, each one tailored to a specific scenario [15]. This scheme has been design to exploit up to three periodical patterns (called seasonalities) in forecasting time-series. The main advantage with respect to standard exponential smoothing methods (ARMA, ARIMA and so on) is the reduced memory requirements and like those, it can be automatically tuned using a training phase.

We selected the single seasonality version in additive configuration, which is a general purpose implementation and it is the most widely used version of the method [10]. As it will be shown in the following, one seasonality has been considered sufficient to analyze the available data-sets. This scheme is based on four equations, reported in Equation 1 for the sake of completeness, that compute the average value of the sequence in the past (Level), the trend of evolution in the future (Trend) and a seasonal term (Seasonality, which allows to exploit the presence of repetitive patterns) and finally the last equation computes the weighted sum of the previous terms as forecast. In the
scheme, $t$ is the time stamp of the input samples ($t = 0, ..., N$, $N$: total number); $k$ is the index of the to-be-forecast samples ($k = 1, ..., K$, $K$: total length of the forecast) and $p$ is the length of the seasonality in samples.

\[
\begin{align*}
    \text{Level : } S_t &= \alpha(X_t - I_{t-p}) + (1 - \alpha) \cdot (S_{t-1} + T_{t-1}) \\
    \text{Trend : } T_t &= \gamma(S_t - S_{t-1}) + (1 - \gamma) \cdot (T_{t-1}) \\
    \text{Seasonality : } I_t &= \delta(X_t - S_t) + (1 - \delta) \cdot I_{t-p} \\
    \text{Forecast : } \hat{X}_t(k) &= S_t + k \cdot T_t + I_{t-p+k}
\end{align*}
\]

**Equation 1.** Additive definition of the Holt-Winters forecasting algorithm.

Now, we show how this scheme perfectly fits data center energy consumption profiles. All the exponential smoothing algorithms require a training phase to tune the parameters ($\alpha$, $\gamma$ and $\delta$) in accordance with input time series’ features. In this case the parameters have been optimized feeding historical data to the scheme and using a multivariate optimization algorithm to explore the state space. Basically the optimization starts with the evaluation of some statistical parameters of the input time-series to initialize the level, trend and seasonality terms, then, iteratively, a multivariate optimization algorithm selects the parameters using a quasi-Newton method (several numerical optimization libraries exist that implement multivariate optimizers like the BFGS algorithm\(^1\)) and with these populates the level, trend and seasonality terms; in the end of each iteration the forecast is evaluated and the error with respect to the reference time-series is computed. The error is then used in the next iteration by the multivariate optimizer to update the set of parameters and the whole process is repeated until the solution converges to the optimum (optimality criterion depends on the numerical optimizer chosen). Figure 2 graphically depicts this optimization process while [10] provides more details about implementation, tuning and optimization methods.

Additionally, we included the prediction update criterion as presented in [8] (named hwt_acp) to recursively refine the HWT forecast by simulating new incoming data in a real-time fashion.

---

\(^1\) The Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm is an iterative method that approximates the Newton’s method for solving unconstrained nonlinear optimization problems.
3.2 DATA-SETS

We used four different data-sets to evaluate and compare the performance of the HWT with the “persistence” method (reference in the following). Figure 3 to Figure 6 show a glimpse of these data-sets, namely from A to D, provided by project partners. In particular, data-sets A, B and C were provided by Credit Suisse and represent the whole consumption of three of their IT rooms, with different workloads. The data-set D instead was provided by Eaton and represents the power consumption of an entire test data center they own. Together these four data-sets illustrate very different scenarios and depict a scenario broader than the urban data center target of the GDN framework, but they resulted very helpful in illustrating the effectiveness of the proposed forecasting approach in the cloud data center scenario as the conclusion of this deliverable will illustrate.

The original time-series (drawn with solid lines and marked as “series” in the pictures) are averaged power consumption samples, acquired every few seconds over a five minutes time-window, corresponding to 288 samples per day. The same series were further averaged to obtain hourly values (drawn with dashed lines and marked as “hourly_avg” in the pictures) and to reduce the number of samples to 24 per day. This choice allowed us to compare the methods both in terms of horizon of
the forecast (one day versus one week) and granularity (288 samples per day as fine-grain case against the coarse-grain, “smoothed”, 24 hourly values per day). All the data-sets span over a two months interval so it was possible to evaluate forecasting horizons of one week at most with a daily seasonality for the HWT seasonality pattern.

**Figure 3.** One week of data from data-set A, solid line marks the original series while dashed line the filtered one.

The main differences among these four traces are the order of magnitude and the deviation from the mean value, specific features that allow characterizing the kind of data center under analysis. Constant and flat power consumption profiles characterize general purpose cloud and corporate data centers, where services reliability must be as close as possible to 100% so all the IT equipment are configured to work with constant performance and hence supply, regardless of the workload and the energy wasted. Rapidly varying power profiles characterize high-performance computing, such kind of systems are located, for example, in research centers where incredible amount of data must be processed in parallel (for example in particle physics and weather forecast tasks) and computing capabilities are offered for lease to anyone can pay. In this case the number of running servers depends on the number of tasks and simulations to execute. For example it may happen with highly parallelized applications where each thread has a lot of data to transfer to/from the storage memory (GBytes), in this case modern servers exploit low power states instead of waiting in idle mode, saving a lot of power and generating the variable consumption profile.
Figure 4. One week of data from data-set B, solid line marks the original series.

Figure 5, shows the power consumption of a small data center consisting of two racks with 5 servers each, and an almost constant workload, this represent a cloud data center of a small company where the maximum variation is lower than 300 VA. The other data-sets, in order two medium- and one large-size data center, show highly variable workloads and power consumption. In particular Figure 4 shows the energy consumption profile of a medium data center targeted to high-performance computing, with last generation servers that can be dynamically switched-on/-off according to the workload. This specific feature can significantly decrease the power consumption of data centers (in this case a variation of approximately 30 kVA), particularly in conjunction with virtualized environments where tasks are virtual machines that can be migrated and packed in the minimum number of physical servers required. Figure 3 and Figure 6 show two cloud data centers where the equipment are less stressed with respect to the high performance computing case. Figure 6 shows also the effect of the installation of new equipment in a preexisting data center, which is an extreme situation that is quite complex to handle with automatic forecasting schemes. Finally we can observe that all of these traces exhibit some periodic patterns that are superimposed to a stable and constant signal. This is the specific characteristic that we are going to exploit by means of the HWT method.
Figure 5. One week of data from data-set C, solid line marks the original series.

Figure 6. One week of data from data-set D, solid line marks the original series.
4. RESULTS

We realized several sets of simulations for each data center; starting with the data-set of 5 minutes span samples we simulated one day and one week horizons (288 and 2016 samples), the same procedure was executed with the hourly averaged data-sets (24 and 168 samples), with all prediction techniques for comparison. Results are presented in terms of MAPE of the forecast with respect to the real trace and are summarized in Figure 7 to Figure 10.

![Figure 7. Comparative results for one day forecast, 5 min sample frequency.](image)

The evaluation of the reference method is pretty straightforward (computed as the difference of two sets of samples corresponding to two consecutive time-horizons) with a quite small memory and data-set size requirements (a trace for one day is enough to predict the following one). The HWT method on the other side, in accordance with the family of exponentially weighted moving averages methods, requires a large data-set only for training the model of the time series to forecast. Once the modeling phase is completed, a number of samples corresponding to the size of the seasonal pattern to exploit is required to compute the prediction. In this case we employed a daily pattern. The data-sets are two months long and with those it was possible to set one week as maximum time horizon in order to have enough data to train the model (six weeks for training, one week as historical data and the last week to validate the prediction and compute the MAPE).
Both methods (hwt_acp is to consider a special case of the Holt-Winters and will be discussed separately) exhibits very high performance for the very short forecast horizon of one day (Figure 7 and Figure 9), with a maximum percentage error lower than 2%. That is remarkably low. In this case we can conclude that using a lower number of samples (hourly averages instead of 5 minutes) does not affect the performance of the forecast but significantly reduces the memory requirements and the execution time of the HWT algorithm (particularly the training phase). Moreover, considering the one day ahead case, we can observe that the proposed approach outperforms the reference method in all the data centers except for the A case, where anyway the error is very small (around 0.3%). This result is particularly of interest since the best performance can be observed in the case of data center B, the one where modern infrastructures and dynamic power management is implemented, as described above. In this case 1% improvement is obtained with samples of 5 minutes span, while almost 0.5% in the hourly case.

On the other hand, considering the weekly horizon of forecast (Figure 8 and Figure 10), clearly the HWT performance are worse than the reference suggesting that for such kind of horizon the daily pattern exploited in the method is not the best option. With larger data-sets it would be possible to evaluate different patterns’ time-length and maybe extend the analysis to include multiple seasonality schemes in the comparison.

Figure 8. Comparative results for one week forecast, 5 min sample frequency.
The hwt_acp refinement scheme further increases the performance of the base HWT one which are, however, really good. Further experiments are required to evaluate its effectiveness since it implies a trade-off between increasing the computational complexity of the forecasting algorithm and improving the prediction (we obtained a gain up to 3% in the weekly forecast with samples of 5 minutes span and data center C).
Finally, Figure 11 and Figure 12, graphical underline the relevance of a forecasting algorithm with respect to the reference persistence approach in case of data-set B and D respectively. In both pictures it is highlighted the error that one could have faced in case a persistence approach had been used with respect to the HWT approach.
Although the percentage error representation, presented before, gives the idea that both forecasting schemes have comparable performance (since averaged on a number of experiments) these two examples underline the impact that a smart approach can have in fulfilling the goals proposed by the GDN framework. With this kind of predictions and in particular its capacity to handle abrupt changes, it would be easier to integrate ancillary services for the smart grid in the resource planning of the whole data center energy management chain of components, for example in the SEMS and AEMS software presented in D3.7 and D3.11 respectively.
5. PORTABILITY

The HWT forecasting software and the other methods evaluated in this deliverable have been implemented in Python (v2.7) for portability and ease of deployment on different platforms. For the sake of the GDN project it will be reshaped in the form of a shared library or standalone tool according to partners’ needs and requirements.

In the current implementation, the software runs within Python’s interpreter command window, and the location of the input time-series can be passed to the hwt command in the invocation, along with other simulation parameters (number of time-slots, seasonality length, etc.).

Input and Output files currently contain comma separated values (CSV) with the following format:

```
Timestamp, Power(kVA)
2014-11-01 01:00:00,18.5001371667
2014-11-01 02:00:00,18.5210731667
2014-11-01 03:00:00,18.5705728333
...
```

where each line represents a time-slot and the corresponding value. Additionally the HWT software produces a log file with all the output information related from the optimization tool, like the Error and the values of the optimized parameters.
6. CONCLUSIONS

“Green” data centers integrate renewable energy sources in their electricity supply chain to reduce the impact of their energy consumption on the electrical Grid. The main issue in this integration process is to optimize the usage of renewable energy by means of energy buffers and dynamically tuning the server’s workload. In this deliverable we present the analysis of two different electricity forecasting methods that are of primary importance to solve the optimization problem. We consider the traditional “persistence” approach against the Holt-Winters exponential smoothing algorithm, that was demonstrated as optimal solution for a very general class of state-space models, has lower memory requirements than the traditional autoregressive methods and has been successfully demonstrated in applications of energy demand forecasting.

Our results show that the proposed approach has better performance than the reference method with highly variable workloads which is particularly of interest for energy-aware “green” data centers.