

# Short-term gas outliers detection using a hybrid ARIMA and Artificial neural network model

Maarten van Someren  
Universiteit van Amsterdam  
Science Park 107  
Amsterdam, The Netherlands  
m.w.vanSomeren@uva.nl

Marco De Nadai  
Universita' degli studi di Trento  
Master's student  
marco.denadai@studenti.unitn.it

## ABSTRACT

The focus of this paper is on building a robust outlier detection and prediction system, which relies on a hybrid Artificial neural network (ANN) and ARIMA model. ARIMA is suitable for linear prediction and ANNs are suitable for non-linear prediction. It is proved that together they can model the complex non-linear relationship between weather forecast variables and the gas consumption. The system labels outliers, which can be reported to the building manager. He can further analyse and fix the HVAC system minimizing in this way the energy waste. The hybrid system was tested with different buildings, predicting the short-term (hourly) gas consumption with a RMSE of 7.995 and detecting all the outliers without the necessity of possessing the previous examples of anomalies.

## Keywords

Energy forecasting, Time series, Artificial neural networks, ARIMA models, outliers, outliers detection, gas consumption prediction, energy forecast

## 1. INTRODUCTION

Energy consumption in buildings is one of the fastest growing sectors. It is measured that the amount of energy consumed in European buildings (households and services) is about 41% of the total energy consumption [12]. Studies and states' directives about minimizing energy consumption and using renewable energy increased steadily with the reduction of fossil fuels, the border frictions with eastern countries like Russia, and the increase of various environmental problems. With this in mind, the European union, with a recent directive [11], has the target to raise EU energy consumption produced from renewable resources to 20%, to reduce by 20% the EU greenhouse gas emissions and to improve by 20% the EU's energy efficiency. This means investments to re-qualify old buildings, new country laws, energy diagnosis, but also new efficiency systems from the used appliances.

Forecasting energy demands has become one of the major research field in the energy departments because it can help gas utilities but also companies and families. Gas utilities buy gas from pipeline companies on a daily bases, so they need to know the needs in advance to be competitive. Companies and families have the aim of reducing the energy consumption and increase efficiency.

Lately, big companies like Google also have shown their interest in this new market, developing thermostats which au-

tomatically control the house climate basing the decisions on the schedule of the users. Nest, a company acquired by Google, declared that customers saved the 11.3% of AC-related energy usage without compromising comfort [29], thanks to the automatic learning implemented in their thermostats. If on one hand the automatic thermostat program setting based on the people behaviour, can help them to save money, anomaly detection can decrease even more the energy consumption. These anomalies usually have a large impact on the final consumption and if caused by deficiency, they represents a huge and easy fixable problem.

In this paper an automatic outlier detection system is proposed, where days/hours with abnormally high and low energy consumption are labelled and reported to the building manager. He can further analyse and fix the HVAC system minimizing the energy waste caused by the outliers. The outlier detection system presented is based on predictions made by a hybrid ARIMA-ANN, which can model linear and non-linear behaviour of the data with very reliable results, and a comparison between the predicted value/trend and the actual one to find outliers.

### 1.1 What is an outlier

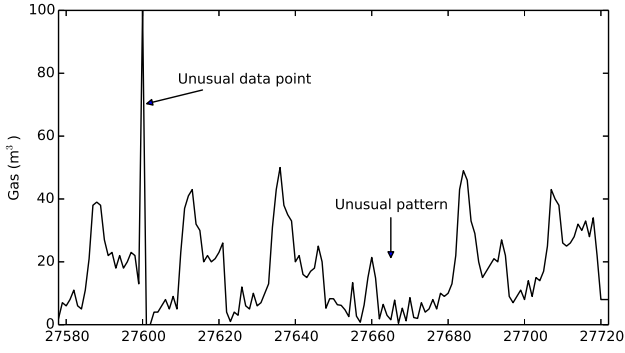
An outlier, by definition [16], is an observation which deviates significantly from other observations so that it creates suspicion that it was created by different dynamics. Outliers can have distinct main reasons:

1. Defective system (e.g. a defective heater in a room).
2. Bad human behaviour (e.g. people who leave open the window in a room while the system is trying to heat it).
3. Defective monitoring system, where the system monitors different values from the real one, due to a malfunction, computing process errors or a recording negligence.

Outliers are also referred to as abnormalities, deviants, novelties or anomalies in the data mining and statistics literature.

As can be seen in fig. 1 there are two main types of outliers unusual data point (e.g. a sudden jump of the energy consumption), or an unusual pattern of changes (e.g. a defective heater that lasts for days). The latter scenario is very challenging and not easily detectable.

## 2. RELATED WORK



**Figure 1: Different types of outliers. On the left an unusual data point is presented, on the right an unusual pattern of changes can be recognized if compared to the other days shape.**

In order to have an excellent anomaly detection, a very accurate prediction model is necessary. This section will review some prediction methods as well as some outlier detection algorithms.

Traditionally, several techniques have been used for energy use forecasting, but it is needed to differentiate between short-term, medium-term and long-term energy forecasting. The former usually refers to prediction with a horizon of hours or days, the second refers to weeks, the latter refers to monthly or annual horizon. Long-term forecasting usually deals with data that rarely presents significant distortions and irregularities, so they have a small effect on the overall value. On the contrary, short-term forecasting has to deal with irregularities and sudden changes of the values (due to weather changes, human behaviour, etc.).

There are essentially five types of prediction models [42]: Engineering methods, Statistical methods, Artificial Neural networks, Support Vector Machines and Grey models. Engineering methods use physical principles to calculate thermal dynamics and energy behaviour of the building, Statistical methods build empirical models to apply a regression to time series of values, Neural networks try to predict energy using an artificial intelligence network of interconnected neurons, Support vector machines are based in a machine learning algorithm and Grey models apply a mixture of the models. All the principal methods are extensively reviewed in [42] and [17].

Several techniques have been traditionally applied for energy use forecasting, and among the statistical methods Kalman filtering and ARIMA/ARMAX time-series techniques are the most famous.

The first reports about applications of Artificial Neural Networks (ANNs) were published in the early 1990's [5]. Since then the number of publications increased steadily. Kalogirou et al. [21] used back propagation neural networks to predict the required heating load of 225 buildings, Ekici and Aksoy used the same model to predict building heating loads in three-buildings. Nizami and Al-Garni [31] tried a simple feed-forward NN and related the electric energy consumption to weather data and population, Taylor and Buizza [38] used an ANN with weather data (51 variables) to predict load of 10 days ahead. Gonzales [15] built an ANN to predict hourly energy consumption. Some re-

searchers tried to specialize the ANNs: Neto and Fiorelli [30] compared generic ANNs with working days ANNs and week-end ANNs, Lazzerini and Rosario [6] specialized them to predict electric lighting with weather data.

Some researchers tried also to apply a hybrid model to increase the performance of the ANN. One example above all is [41] which applied a hybrid ARIMA and neural network model to forecast electricity use, another one is [22] who improved the previous one. This paper is based also on his work.

Until now only electric forecasting was presented, because the majority of the existing forecasters are related to electric forecasting. There are only few of them are about natural gas forecasting: Brown et al. [3] built one of the first predictor for natural gas consumption and Khotanzad et al. [23] developed a two stage system ANN with very good results.

Even if ANNs might outperform traditional methods, the researchers are not still convinced about the results of the ANNs in this field. Nevertheless it is also stated that "a significant portion of the NN research in forecasting and prediction, lacks validity" [1] and that most of the papers seems misspecified models that had been incompletely tested (no standard benchmarks, no synthetic data, etc.) [17]. This paper will try to avoid these mistakes.

Outlier detection system is a wide area from introduction detection systems to fraud detection systems, from law enforcement systems to earth science anomaly detection systems. Talking about energy consumption outlier detections, usually literature is based on the Gaussian error theory, stating that when the measurement accord with normal distribution, the probability that the residual falls in three times the variance is more than 99.7%. Therefore the residuals falling outside it, can be considered outliers. In [13] the author further improved this system considering a rolling moving window median which seems to improve the results when the distribution is not fixed. Other outlier detection mechanisms are based on supervised classifications using trees, ANNs and other different algorithms, thanks to the possession of previous examples of anomalies. The last category of outliers detection system is represented by unsupervised methods like clustering, where an algorithm tries to find similarities between points/trends and cluster them into groups, calculating the distance between them. A cluster is good when the intra-cluster distance is minimized and the intra-cluster distance is maximized. Popular methods in this group are *k*-means, one-class SVM and self-organizing maps. They are very difficult to apply in time-series and the results are not usually excellent after it. For this reason they not considered in this paper.

### 3. PROPOSED SOLUTION

It can be seen that in academic and industry research, linear regression-based systems are the standard "de facto" of energy forecasting and in recent works this problem is treated by combining weather forecast data. However, this relationship is clearly non-linear [17]. Consequently, even if some papers have good results with measured datasets, these systems cannot adequately capture the relationship in all the situations and data. Since ANNs are the state of the art technique of many machine learning problems where there are complex non-linear hypothesis, the proposed solution is composed by a *multilayer feed-forward* neural network with *backpropagation*. This model is applied to the dataset and

**Table 1: Buildings used**

Building name	Date interval	Number of rows
HvA 740 - NTH	01/2008 - 03/2014	54.725
Hva 761 - KMH	01/2008 - 03/2014	54.744
Hva 882 - WBW	01/2008 - 03/2014	54.647

the outlier detection will be built on top of it. All the points far from the predicted value of the ANN, will be considered outliers.

### 3.1 Experimental data

The energy consumption datasets used are collected by Ebatech<sup>1</sup> in different buildings of the Hogeschool van Amsterdam<sup>2</sup>. These buildings are located in Amsterdam, the capital city of The Netherlands. This city has an oceanic climate, strongly influenced by the North Sea. Winters are cold and summers are rarely hot. It is characterized by the common presence of rain and wind.

Ebatech collected different types of features in each buildings, with different granularity. For this project three buildings are used: *HvA 740 - NTH*, *Hva 882 - WBW* and *Hva 761 - KMH*. In these buildings the company collected the energy consumptions, the gas consumptions as other variables.

The weather data was collected by KNMI<sup>3</sup> in Schipol, a Dutch city around 16 km far from the tested buildings. The available data consists in over 21 variables hourly collected. The proposed solution only uses few of them, as explained in the section 3.1.2, and they are used as forecast values: the measured weather conditions are linked to the previous hour of energy consumptions. It is necessary to consider that there will be an error in the built model, since the weather data is collected in a different location from the buildings positions and in practice the error will be larger than those obtained in this simulation, due to the effect of the weather forecast uncertainty [9, 34]. The advice is to keep it in mind before applying the methods contained in this paper with days forecasting.

#### 3.1.1 Data analysis

The available dataset is conform to the nine characteristics of a good dataset as outlined by to Few [14]:

**Volume** As demonstrated in table 1, the dataset seems adequately large to train an ANN and test the results.

**Historical** The dataset covers a very large period ranging more than 5 years, allowing to see similar patterns even with different yearly/seasons behaviour (one year could be different from another one for external factors like weather or building use).

**Consistent** The dataset of the selected buildings and for the variables that will be used, is consistent.

**Multivariate** The dataset is composed by more than 50 variables, measuring the different behaviours of the buildings and the weather.

<sup>1</sup><http://www.ebatech.nl>

<sup>2</sup><http://www.hva.nl/>

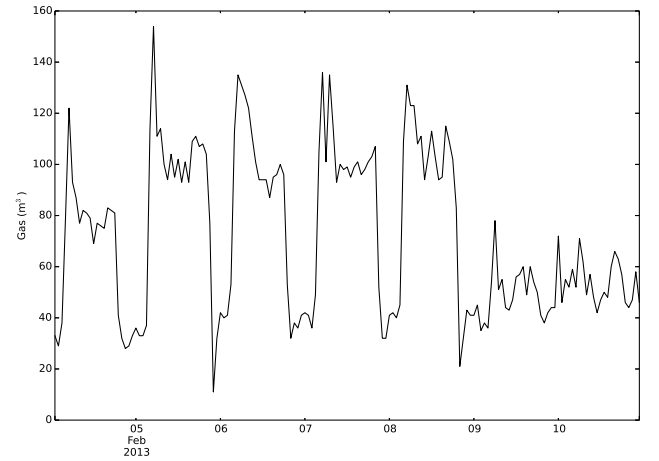
<sup>3</sup>Koninklijk Nederlands Meteorologisch Instituut <http://www.knmi.nl>

**Clean-Atomic** The dataset, after a minimal preprocessing, is clean and atomic. Some errors (outliers) were present on the data, but they were selected and deleted before the use.

**Clear** The dataset of the selected buildings and for the variables that will be used, is clear. The only problems were found on the not used variables: although they could be very useful, it was not possible to contact the buildings energy managers in order to understand them, for this reason they were excluded.

**Richy Segmented** The dataset can be segmented in several meaningful segments like gas consumption, electric consumption, water consumption, number of people using the buildings, etc.

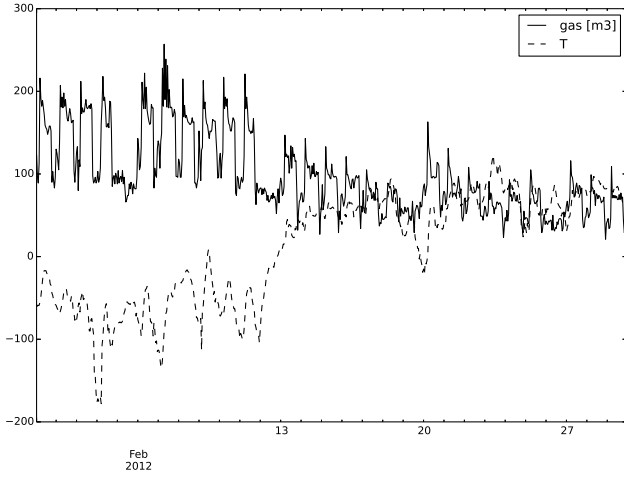
The gas consumption data is highly seasonal: daily and weekly cycles are quite perceptible, as it can be seen from fig. 3 and fig. 2. From the latter the weekly behaviour is clear: the last two days of the week (Saturday and Sunday) are completely different from the others and Monday seems a bit different from the rest of the days. Everyday, around 4:00-5:00 AM, the system seems to react turning on the heating system, where in the previous hours of the night it seems only keeping a minimum temperature. The system reveals us that after a couple of hours, it decreases the consumption again. In fig. 3 the Temperature has a clear daily/hourly relation with the gas consumption while in fig. 4 the electric consumption is shown to be very smoothed and more regular than the gas one.



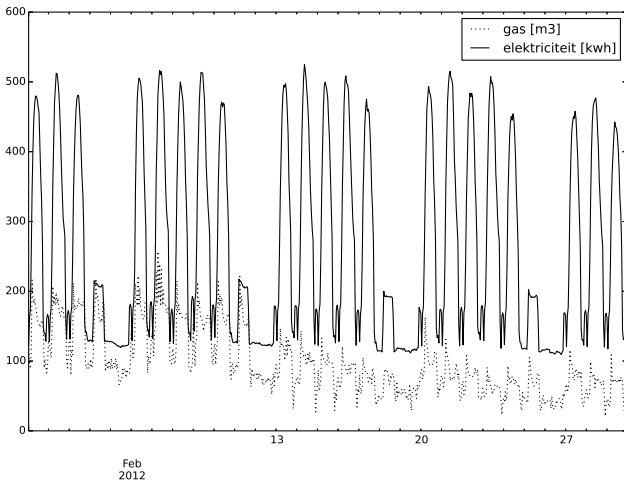
**Figure 2: Typical weekly and daily gas consumption behaviour. The weekly pattern can be noticed by observing that the last two days of the week (Saturday and Sunday) have a completely different shape than the others. During the week, the daily behaviour is very similar, with one peak around 4:00-5:00 AM.**

#### 3.1.2 Data preprocessing

Since NNs are sensitive to missing values and irregularities, some data considered evident outliers will be removed, nevertheless it needs to be considered that since it was not possible to contact the buildings manager, it was not possible to remove other outliers. For this reason the ANN



**Figure 3: Typical monthly gas consumption behaviour and its relation with the temperature**



**Figure 4: Relation between electric and gas consumption**

training was done with not entirely perfect data; having the possibly to further clean the data could improve the final results.

### 3.2 Artificial neural network predictor

The predictor is a *Multilayer Feedforward* ANN with a feedback structure, called *Backpropagation*, which calculates the error in the output layer and then back-propagates it to the previous layers in order to adjust the weights [35]. The ANN is a 1-hidden-layer network composed by neurons with a sigmoid transfer function as suggested in [26], and one output node. Training is characterized by a learning rate of 0.001 and fixed by a Momentum of 0.01, which could help to increase the speed, avoiding local minima.

The available dataset is essentially a time-series, a vector of scalars which depend on time  $t$ :

$$\{x(t_0), x(t_1), \dots, x(t_i), x(t_{i+1}), \dots\}$$

where the estimation of  $x$  at some future time with a  $s$  time

horizon, is needed.

$$\hat{x}(t+s) = f(x(t), x(t-1), \dots)$$

As a result the ANN needs to predict the next value in dependence of the “acquired knowledge” which comes from the actual values and during the previous training. This is realized by a moving window containing a “memory” of the previous states for the interesting variables, using a *Tapped delay line memory* [28]. These memories form a new set of states  $\{\bar{x}_1(t), \bar{x}_2(t), \dots, \bar{x}_n(t)\}$ , from the original states  $\{x(1), x(2), \dots, x(n)\}$ , where  $\bar{x}_i(t) = x(t-i+1)$ . The windows will be explained later.

Since the value to predict is time dependant, the first element to consider is to add the time as a feature. Energy consumption depends on the hour of the day but also on the day of the week (as explained in section 3.1.1, fig. 3 and fig. 2). The hour is coded by means of its sine and cosine values as usual reported in literature [33, 8, 15]. The days were firstly coded as week days (from 0 to 6, where 0 is Monday and 6 is Sunday), then coded by their sine and cosine as for the hours. Since the behaviour of the holidays were considered similar to the week-ends (particularly similar to Sundays), a function encoded all the holidays as week-end days<sup>4</sup>. Regardless of whenever the school was operational, the first tests showed that the Ebatech system was normally heating the buildings (Christmas, in Tuesday 25th of December 2012, was heated like a normal Tuesday even if the building was certainly closed). This causes an avoidable waste, which was immediately reported to the company. For this reason, this function was not applied. Energy consumption is highly seasonal and since the predictor is learning from years of data, the day of the year (from 0 to 365) is included as a feature, transformed by its sine and cosine. The third element added is the current system load, which is the energy consumption at the  $k$  state when the load at  $k+1$  needs to be predicted. Although the electric consumption is almost constant, it was found to be somehow helpful to predict the gas consumption, so it was considered a feature.

There are many factors that affect the energy needs of buildings. These factors can be divided in three main groups namely the physical environmental, the artificial designing parameters and the human thermal discomfort. The first is composed by weather related parameters like outdoor temperature, wind speed, solar radiation, etc. The artificial designing parameters are related to the building construction: transparency ratio, orientation, etc [10], but these variables were unavailable in the dataset. The human sensation of thermal discomfort is correlated not only to the temperature, but also to other variables such as relative humidity, irradiation and wind speed. Even if all these data was available, the only significant weather variables found, were the temperature and the wind speed.

The system consumption was believed to be related to the difference of the outdoor temperature between two instants (eq. (1)), representing a positive/negative change of the external environmental conditions.

$$\Delta T_{k+1} = T_{k+1} - T_k \quad (1)$$

where  $T_{k+1}$  is the predicted temperature for the period  $k+1$  and  $T_k$  is the value measured in the instant  $k$ . It needs to be

<sup>4</sup>Thanks to an Open source dutch week-end list <https://github.com/PanderMusubi/dutch-holidays>

noted that the real behaviour of the system was unknown, so it was not possible to know if this change would have an immediate effect on the HVAC system and/or its reaction time.

As pointed out by Zhang et al [40], ANNs models really have advantages while dealing with a large amount of historical load data with non-linear characteristic, but the researchers neglected the linear relations including the data. For this reason a hybrid approach is proposed, where the ANN will be helped in linear forecasting by the popular method ARIMA (autoregressive integrated moving averages), commonly known as the BoxJenkins approach. Stationarity is a necessary condition in building an ARIMA model used for forecasting. A stationary time series is characterized by statistical characteristics such as the mean and the autocorrelation structure being constant over time. Consecutively, the data was de-trended and the variance was stabilized. In order to do this, the time series was analysed by the STL decomposition by LOESS [4], which decomposes a time series into seasonal, trend and irregular components by an additive method. The trend is a long-term increase or decrease of the data. The seasonal part is the time series affected by seasonal factors. The irregular components are the remainder of the original time series without the seasonal factors, trend and cycles. Gas usage has a daily cycle but there are also secondary weekly and annual cycles that ARIMA may not be able to capture. As stated before, there is also a very noticeable difference between the weekdays and the week-end data. For this reason, the time series was firstly analysed to extract the yearly seasonal part (fig. 5), then studied to discard the daily seasonal part (since STL does not support more than one seasonal part)<sup>5</sup>. To further help ARIMA, a subset of the data was created, splitting it into weekdays and weekend days. The ARIMA model was separately applied to them.

Several diagnostic tests were applied to the resulting time series, considering a significance level of 5%: the Ljung-Box test [27], The Augmented DickeyFuller (ADF) t-statistic test [7] and The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test [24]. The most suitable model found was ARIMA(3,0,5) and it was applied to forecast the gas consumption one hour ahead. The first tests proved that ARIMA helped the ANN model, but a further test showed that including the STL yearly seasonal residuals and the STL weekly seasonal residual helped even more the neurons to forecast the gas consumption. For this reason they were included as well in the input set. An ARIMA model with temperature dummy variables was tested, but it was not helping further the ANN. Since the simplest models are always preferred, this one was discarded.

Taking into account the points made in this section, the ANN is predicting the gas consumption “seeing” without knowing its shape, and its behaviour in the previous hours/days. This limit is surpassed by some rolling windows, which will somehow simulate the Recurrent neural networks behaviour. Two rolling windows were created for the gas consumption, memorizing the sum and the peak load of the previous five hours, and two moving rolling windows were created for the STL yearly residuals, memorizing sum and peak of them.

All the features that were not between the limits  $[-1, 1]$

<sup>5</sup>Fourier terms were tested (as explained in <http://robjhyndman.com/hyndsight/dailydata/>) but they were not helpful more than STL.

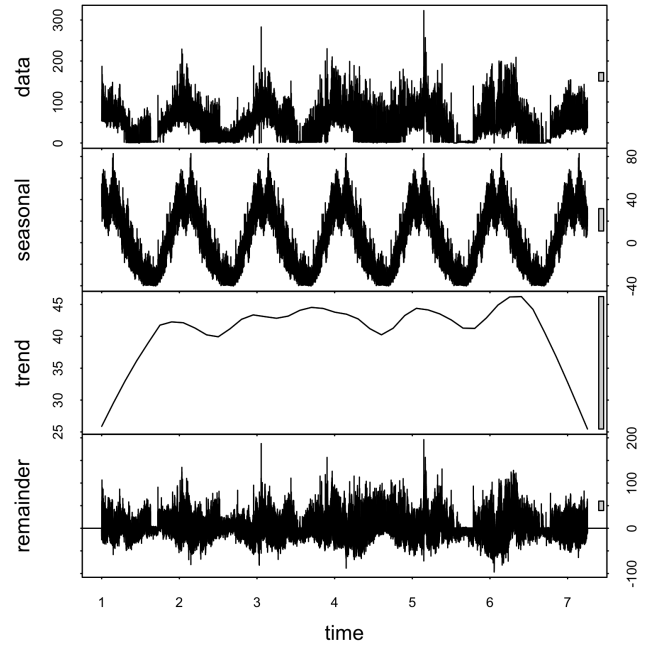


Figure 5: Yearly STL decomposition by LOESS

were scaled to have a faster convergence of the *Stochastic Backpropagation* [26].

$$x'_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (2)$$

where  $x_i$  is the original value and  $x'_i$  is the scaled one.

Choosing number of hidden units for the ANN is always a tricky task. As stated by [25, 36], using early stopping in a oversized *Backpropagation* ANN, where the number of hidden neurons is higher than the number of the features, makes easier to find global optimum and avoid bad local optima. For this reason the number of hidden units was chosen to be greater than  $2 \times |\text{features}|$  and then test-driven.

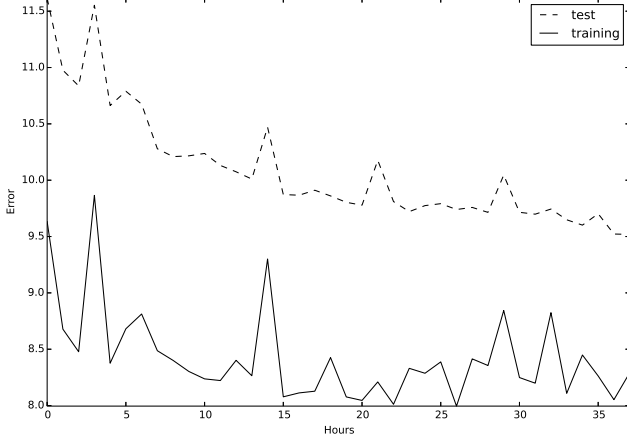
This project used Python and Pybrain [37] to construct the planned ANN and the R forecast package of Rob J.

Table 2: ANN inputs

Variable	Data
Electricity load	$E(t)$
Hour	$\sin(2\pi(h)/24); \cos(2\pi(h)/24)$
Week day	$\sin(2\pi(wDay)/6); \cos(2\pi(wDay)/6)$
Year day <sup>6</sup>	$\sin(2\pi(d)/365.25); \cos(2\pi(d)/365.25)$
Gas peak load'	$\max_{1 \leq k \leq 5} G(t-k)$
Gas sum load'	$\sum_{i=1}^5 G(t-i)$
Gas mean load'	$\frac{1}{288} \sum_{i=1}^{288} G(t-i)$
Temperature	$T(t)$
Wind speed	$FH(t)$
$\Delta T_{k+1}$	$T_{k+1} - T_k$
ARIMA forecast	$\text{forecast}(\text{ARIMA}(3, 0, 5), 1)$
STL year res.	$\text{YearRes}(t)$
STL day res.	$\text{DayRes}(t)$
Year res. peak	$\max_{1 \leq k \leq 5} \text{YearRes}(t-k)$
Year res. sum	$\sum_{i=1}^5 \text{YearRes}(t-k)$

**Table 3: ANN best selected results to compare the hybrid model with the ANN one. This shows that the Hybrid model outperform the ANN one by 66%.**

Model	neurons	epochs <sup>7</sup>	RMSE	MAPE	MAE
ANN	80	18	11.95	30.01	7.52
Hybrid	60	12	8.21	27.14	5.57
Hybrid	80	26	7.99	27.62	5.53



**Figure 6: Training curve of the hybrid model with 80 hidden neurons**

Hyndman [19].

### 3.3 Outlier detection

According to Chebschev's theorem [2], almost all the observations in a data set of system states falls into the interval  $[\mu - 3\sigma, \mu + 3\sigma]$ , where  $\mu$  and  $\sigma$  are respectively the mean and standard deviation of the data set, and the data points outside this interval are declared outliers. In this paper the ANN is used to predict the gas consumption, for this reason a point will be considered outlier if it will fall outside the 95% confidence interval<sup>8</sup> expressed for the RMSE. If it is assumed that the difference between the actual values  $x_i$  and the predicted value  $\hat{x}_i$  have:

$$\hat{x}_i - x_i \sim \mathcal{N}(0, \sigma^2) \quad (3)$$

- mean zero.
- follow a Normal distribution (it is assumed that it holds for the large amount of data utilized).
- and all have the same standard deviation  $\sigma$ .

$$\hat{x}_i - x_i \sim \mathcal{N}(0, \sigma^2) \quad (4)$$

it is possible to say that eq. (4) follows a  $\chi_n^2$  distribution

<sup>8</sup>For the following description user *fabee* of CrossValidated.com needs to be cited <https://stats.stackexchange.com/questions/78079/confidence-interval-of-rmse>

with  $n$  degrees of freedom. Which means:

$$P\left(\chi_{\frac{\alpha}{2}, n}^2 \leq \frac{n\text{RMSE}^2}{\sigma^2} \leq \chi_{1-\frac{\alpha}{2}, n}^2\right) = 1 - \alpha \quad (5)$$

$$\Leftrightarrow P\left(\frac{n\text{RMSE}^2}{\chi_{1-\frac{\alpha}{2}, n}^2} \leq \sigma^2 \leq \frac{n\text{RMSE}^2}{\chi_{\frac{\alpha}{2}, n}^2}\right) = 1 - \alpha \quad (6)$$

$$\Leftrightarrow P\left(\sqrt{\frac{n}{\chi_{1-\frac{\alpha}{2}, n}^2}}\text{RMSE} \leq \sigma \leq \sqrt{\frac{n}{\chi_{\frac{\alpha}{2}, n}^2}}\text{RMSE}\right) = 1 - \alpha. \quad (7)$$

Therefore

$$\left[\sqrt{\frac{n}{\chi_{1-\frac{\alpha}{2}, n}^2}}\text{RMSE}, \sqrt{\frac{n}{\chi_{\frac{\alpha}{2}, n}^2}}\text{RMSE}\right] \quad (8)$$

## 4. EXPERIMENTAL EVALUATION

The ANN has been trained with early stopping, with a fixed number of training epochs (phases) or stopping the training when the validation error rate was increasing. All the results showed are obtained from a k-fold cross-validation techniques, where the network was trained  $k$  times, each time leaving out a subset of data from training in order to test the ANN. The results of the  $k$  tests, were divided by  $k$ .

Although most the Mean Absolute Percentage Error (MAPE) is considered a standard for examining the quality of the models prediction of energy load, it is an adequate error measure only if the loss function were linear and recent studies demonstrated it is not [21][20]. Moreover the percentage error is infinite if there are zero values on the series, frequent in intermittent data and in consumption data, and it puts a heavier penalty on positive errors than on negative errors [18]. Because of these disadvantages, this paper will only consider the minimization of the Root Mean Square Error (RMSE), which penalize large errors, as suggested in [39]. As suggested in [17], for every experiment also the Mean Absolute Error (MAE) will be calculated.

$$\text{MAPE}^9 = \frac{1}{n} \sum_{i=1}^n \frac{|Y_i - \hat{Y}_i|}{Y_i} \times 100 \quad (9)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2} \quad (10)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{Y}_i - Y_i| \quad (11)$$

where  $\hat{Y}$  is the vector of the  $n$  predictions and  $Y$  is the vector of the true values.

### 4.1 Synthetic experiments

The method is tested with synthetic generated data. Two days outliers were generated by different algorithms. In the first one the real consumption was modified by a random

<sup>9</sup>MAPE errors will be calculated only on the non-zero values, to avoid the problems described before.

value, simulating the system measurement/control malfunction which makes the consumption bouncing up and down (see eq. (12)). The second synthetically created day was created adding  $50m^3$  of gas consumption to the real one, creating a pattern which simulates a strange behaviour and/or a malfunction of the heating system (see eq. (13)).

$$G(t) = G(t) + v * 30 \quad (12)$$

where

$$v \sim \mathcal{N}(0, \sigma^2)$$

$$G'(t) = G'(t) + 50 \quad (13)$$

The two outliers were correctly detected, as it can be seen in fig. 7.

## 4.2 Measured data experiments

The method is also tested with measured data, coming from a different type of the day. For example the gas consumption of a week-end was placed in a week-day, simulating a holiday. The purpose of this test was to show that the unusual pattern was detected. In fig. 8 it can be seen that the outlier mechanism works perfectly when the Sunday gas consumption is placed in a weekday.

The outlier was correctly detected, as it can be seen in fig. 8. The robustness of the design was proved with different building, listed in table 1.

## 5. CONCLUSION

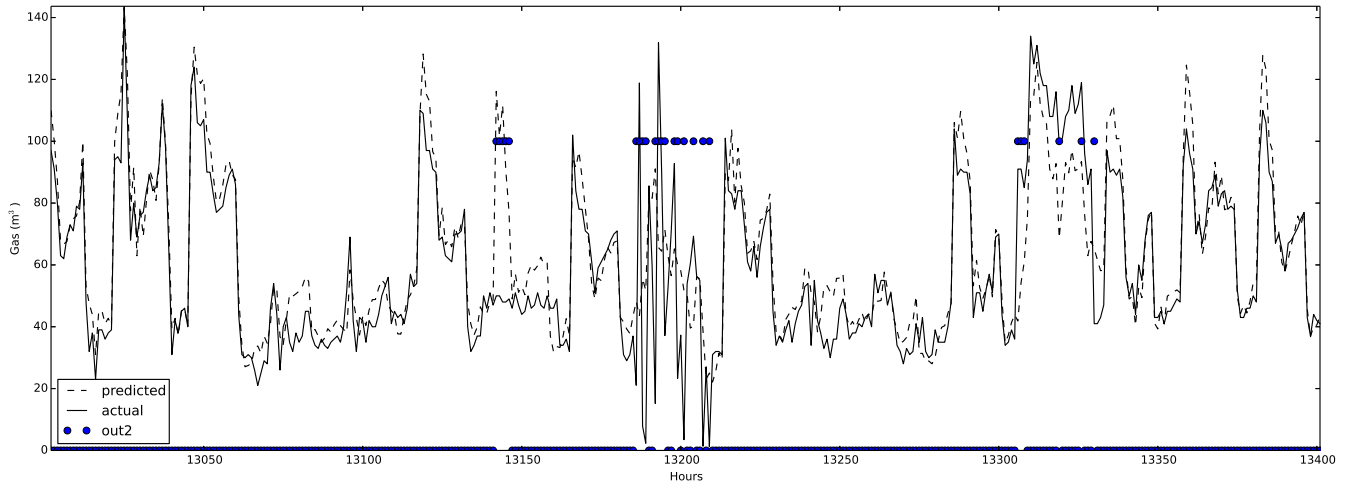
Any model can't be treating accurately all the situations for a large amount of historical load data. The irregular fluctuation of the gas consumption was hardly predictable, for this reason the ANN model needed help from a classic linear model like ARIMA. Although other papers presented similar models to forecast electric consumption, the hybrid model presented here, is almost unique because it focuses to forecast short-term gas consumption, which are very irregular and not easily predictable with classic methods. Since the predictor is very accurate (with RMSE of 7.995), the outlier mechanism is able to easily detect strange behaviours defining a threshold value in the confidence interval without the need to possess previous example of outliers.

## 6. ACKNOWLEDGMENTS

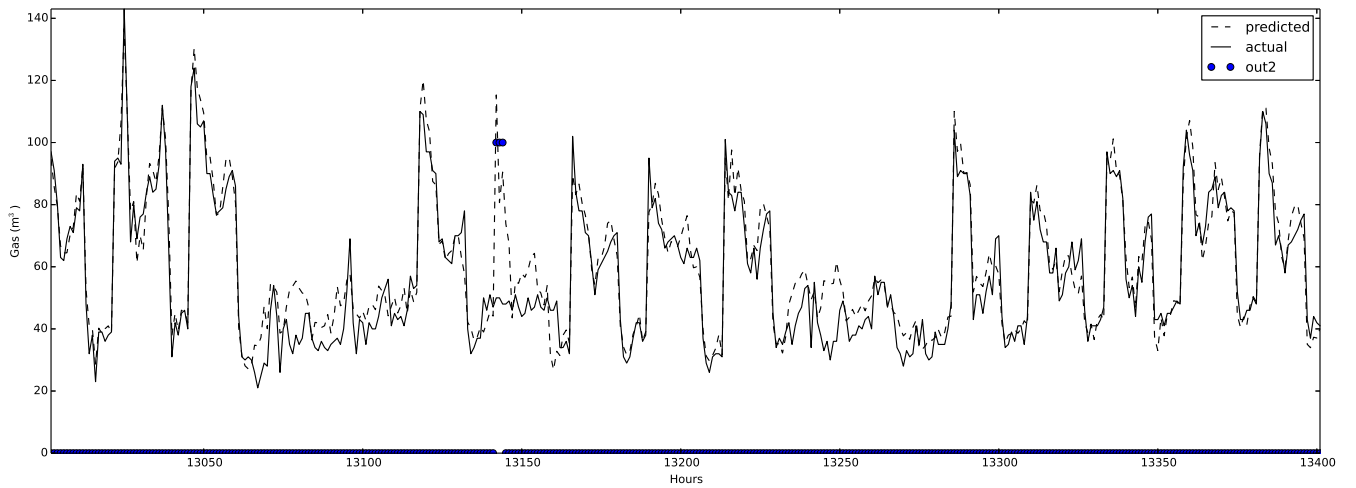
This work and internship was economically supported by Università degli studi di Trento.

## 7. REFERENCES

- [1] M. Adya and F. Collopy. How effective are neural networks at forecasting and prediction? a review and evaluation. *J. Forecasting*, 17:481–495, 1998.
- [2] B. G. Amidan, T. A. Ferryman, and S. K. Cooley. Data outlier detection using the chebyshev theorem. In *Aerospace Conference, 2005 IEEE*, pages 3814–3819. IEEE, 2005.
- [3] R. H. Brown and I. Matin. Development of artificial neural network models to predict daily gas consumption. In *Industrial Electronics, Control, and Instrumentation, 1995., Proceedings of the 1995 IEEE IECON 21st International Conference on*, volume 2, pages 1389–1394. IEEE, 1995.
- [4] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990.
- [5] T. Czernichow, A. Piras, K. Imhof, P. Caire, Y. Jaccard, B. Dorizzi, and A. Germond. Short term electrical load forecasting with artificial neural networks. *Engineering Intelligent Systems for Electrical Engineering and Communications*, 4(LRE-ARTICLE-1996-003):85–99, 1996.
- [6] E. D'Andrea, B. Lazzerini, and S. L. del Rosario. Neural network-based forecasting of energy consumption due to electric lighting in office buildings. In *Sustainable Internet and ICT for Sustainability (SustainIT), 2012*, pages 1–5. IEEE, 2012.
- [7] D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431, 1979.
- [8] R. H. Dodier and G. P. Henze. Statistical analysis of neural networks as applied to building energy prediction. *Journal of solar energy engineering*, 126(1):592–600, 2004.
- [9] A. P. Douglas, A. M. Breipohl, F. N. Lee, and R. Adapa. The impacts of temperature forecast uncertainty on bayesian load forecasting. *Power Systems, IEEE Transactions on*, 13(4):1507–1513, 1998.
- [10] B. B. Ekici and U. T. Aksoy. Prediction of building energy consumption by using artificial neural networks. *Advances in Engineering Software*, 40(5):356–362, 2009.
- [11] European Parliament and Council of the European Union. Directive 2009/28/EC, 2009.
- [12] E. C. Eurostat. *Energy balance sheets - 2010-2011 - 2013 edition*. Publications Office of the European Union, Luxembourg, 2013.
- [13] H. Ferdowsi, S. Jagannathan, and M. Zawodniok. A neural network based outlier identification and removal scheme. In *Prognostics and Health Management (PHM), 2013 IEEE Conference on*, pages 1–6. IEEE, 2013.
- [14] S. Few. *Now you see it: simple visualization techniques for quantitative analysis*. Analytics Press, 2009.
- [15] P. A. González and J. M. Zamarreño. Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. *Energy and Buildings*, 37(6):595–601, 2005.
- [16] D. Hawkins. Identification of outliers. *London: Chap*, 1980.
- [17] H. S. Hippert, C. E. Pedreira, and R. C. Souza. Neural networks for short-term load forecasting: A review and evaluation. *Power Systems, IEEE Transactions on*, 16(1):44–55, 2001.
- [18] R. J. Hyndman. Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, 4(4):43–46, 2006.
- [19] R. J. Hyndman and Y. Khandakar. Automatic time series for forecasting: the forecast package for r. Technical report, Monash University, Department of Econometrics and Business Statistics, 2007.



**Figure 7: Outlier detection with synthetic generated data. The circle represents the hours where an outlier is detected.**



**Figure 8: Outlier detection where the gas consumption of a weekday was replaced by a Sunday one. The (three) circles represent the outliers detected by the system.**



- [20] S. Kaji, M. Roberge, L. Lamarche, and P. Malinowski. Evaluation of building energy consumption based on fuzzy logic and neural networks applications. In *Proc of CLIMA*, page 264, 2000.
- [21] S. A. Kalogirou. Artificial neural networks in energy applications in buildings. *International Journal of Low-Carbon Technologies*, 1(3):201–216, 2006.
- [22] M. Khashei and M. Bijari. An artificial neural network ( $i_1, p_1, i_2, d_1, i_3, q_1, i_4$ ) model for timeseries forecasting. *Expert Systems with Applications*, 37(1):479–489, 2010.
- [23] A. Khotanzad, H. Elragal, and T.-L. Lu. Combination of artificial neural-network forecasters for prediction of natural gas consumption. *Neural Networks, IEEE Transactions on*, 11(2):464–473, 2000.
- [24] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54(1):159–178, 1992.
- [25] S. Lawrence, C. L. Giles, and A. C. Tsoi. What size neural network gives optimal generalization? convergence properties of backpropagation. 1998.
- [26] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.
- [27] G. M. Ljung and G. E. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.
- [28] M. C. Mozer. Neural net architectures for temporal sequence processing. 2007.
- [29] Nest. Energy savings from nest white paper preview. <https://nest.com/downloads/press/documents/efficiency-simulation-white-paper.pdf>, 2014.
- [30] A. H. Neto and F. A. S. Fiorelli. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy and Buildings*, 40(12):2169–2176, 2008.
- [31] S. J. Nizami and A. Z. Al-Garni. Forecasting electric energy consumption using neural networks. *Energy Policy*, 23(12):1097 – 1104, 1995.
- [32] O. U. of Catalonia. The importance of being (r)eproducible. <http://labs.biblioteca.uoc.edu/blog/?p=4332>, 2014.
- [33] M. B. Ohlsson, C. O. Peterson, H. Pi, T. S. Rognvaldsson, and B. P. Soderberg. Predicting system loads with artificial neural networks—methods and results from” the great energy predictor shootout”. *ASHRAE Transactions-American Society of Heating Refrigerating Airconditioning Engin*, 100(2):1063–1074, 1994.
- [34] D. K. Ranaweera, G. G. Karady, and R. G. Farmer. Effect of probabilistic inputs on neural network-based electric load forecasting. *Neural Networks, IEEE Transactions on*, 7(6):1528–1532, 1996.
- [35] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.
- [36] W. S. Sarle. Stopped training and other remedies for overfitting. In *Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics (.’. -\.* pp. 352–360. Interface Foundation of North America, Fairfax Station. VA, USA, 1995.
- [37] T. Schaul, J. Bayer, D. Wierstra, Y. Sun, M. Felder, F. Sehnke, T. Rückstieß, and J. Schmidhuber. PyBrain. *Journal of Machine Learning Research*, 2010.
- [38] J. W. Taylor and R. Buizza. Neural network load forecasting with weather ensemble predictions. *Power Systems, IEEE Transactions on*, 17(3):626–632, 2002.
- [39] R. Yao and K. Steemers. A method of formulating energy load profile for domestic buildings in the uk. *Energy and Buildings*, 37(6):663–671, 2005.
- [40] G. Zhang, B. Eddy Patuwo, and M. Y Hu. Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting*, 14(1):35–62, 1998.
- [41] G. P. Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.
- [42] H.-x. Zhao and F. Magoulès. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16(6):3586–3592, 2012.

## APPENDIX

### A. SOURCES

“Very often scientific studies rely on complex textual explanations of what has been done to analyse the data that can overwhelm the reader that has to accept them as an act of faith”, as stated in [32]. To avoid this, the author provides all the sources through this website: <https://github.com/denadai2/energyUva>. The raw data are protected by privacy of the Hogeschool van Amsterdam. The Universiteit van Amsterdam is working to make them available to achieve total reproducibility.