Short-term load forecasting using a Gaussian process model: The influence of a derivative term in the input regressor

J.M. Lourenço\textsuperscript{a,\ast} and P.J. Santos\textsuperscript{b}

\textsuperscript{a}Department of Systems and Informatics, Polytechnic Institute of Setúbal, INESC-ID, Setúbal, Portugal
\textsuperscript{b}Department of Electrical Engineering, Polytechnic Institute of Setúbal, INESC-Coimbra, Setúbal, Portugal

Abstract. This paper establishes a regressive methodology, applied to the electrical distribution sector, to predict the next hour load. The approach followed is based on a Gaussian process model. To create a simple structure for the input regressor only a few instances of the endogenous variables, based on the homologous and contiguous values, were included. Information that allows to evaluate the influence of a derivative term in the quality of the forecasts is also included in the input vector. In order to assess the performance of different load situations the model was tested on a real-load case-study. The case study includes three different electrical distribution substations, representative of typical load consumer's patterns, namely the residential, non-residential and the service sector. The results obtained are in accordance with the values normally accepted in the sector.

Keywords: Gaussian process models, load forecasting, distribution systems, electrical distribution substations

1. Introduction

Distribution companies operating in a scenario of complete or partial unbundling of the electricity sector are confronted with increasing demands on planning, management and operation of the networks. Relations with generation, transmission and retail companies are now becoming increasingly complex \[9\]. Therefore, the distribution companies play a major role in the managing and planning of distribution, with an emphasis on the quality of the supply. System operators have to use as much as possible reliable data, namely on load forecast results, having in mind simultaneously that uncertainty is a key issue for most decisions \[12\].

Short-Term Load Forecasting (STLF) methodologies have registered an evolution during the last four decades \[3\] starting with approaches based on time series analysis through statistical methods into knowledge-based systems, and more recently, other approaches based on fuzzy algorithms, artificial neural networks and genetic algorithms \[4,10\]. The diffuse relations between load and other relevant variables and lack of information have led to a clear development of these most recent methodologies that are recognized to deal better with uncertainty \[14\].

In this paper we develop a method for short-term forecasting based on Gaussian process (GP) models. These are probabilistic non-parametric models that provide the prediction and the variance of the prediction \[16\]. The diffuse relations between load and other relevant variables and lack of information have led to a clear development of these most recent methodologies that are recognized to deal better with uncertainty \[14\].

In this paper we develop a method for short-term forecasting based on Gaussian process (GP) models. These are probabilistic non-parametric models that provide the prediction and the variance of the prediction \[16\]. The diffuse relations between load and other relevant variables and lack of information have led to a clear development of these most recent methodologies that are recognized to deal better with uncertainty \[14\].

Another advantage of this kind of models is that its structure is determined only by the selection of the covariance function and the regressors that can be determined even with relatively small training sets. The
number of tuning parameters of the covariance function is smaller than the number of parameters of a conventional neural network. The disadvantage of the method is the potential computational load associated with the inversion of the covariance matrix, which increases with the amount of data and number of regressors, for optimization and prediction.

The methodology proposed is illustrated by a case study (Fig. 1) composed of three medium-size supply electrical distribution substations (ES) [15]. The three ES are supplied at 60 kV and the secondary voltage is 15 kV, a voltage level that is normal in this type of distribution network.

To test the model under different load conditions three different ES were considered. These ES have different types of load patterns, namely the residential, non-residential, and a third one representative of the service sector (Fig. 2). This last pattern has two typical zones with two peak loads, one in the morning and the other in the evening, that present an increased difficulty to the forecast model because of the unknown relative position of the two peak loads (morning and evening).

2. Description of the Gaussian process model

A Gaussian process is a collection of random variables, which have a joint multivariate Gaussian distribution. Assuming a relationship of the form \( y = f(x) \) between an input \( x \in \mathbb{R}^D \) and an output \( y \in \mathbb{R} \), the output can be viewed as a collection of random variables \( y(1), \ldots, y(n) \sim \mathcal{N}(0, \Sigma) \) that have a joint multivariate Gaussian distribution. The covariance matrix \( \Sigma \) can be parameterized and computed by means of a function \( \sum_{pq} = \text{Cov}(y(p), y(q)) = C(x(p), x(q)) \) that determines the covariance between output points corresponding to input points \( x(p) \) and \( x(q) \). The Gaussian process can be fully specified by its mean \( \mu(x) \) (usually assumed to be zero) and its covariance function \( C(x(p), x(q)) \). It is remarked that, although not all data can be modeled as a zero-mean process, this assumption is correct if the data is properly scaled and detrended [7].

A common choice of covariance function, when we assume that the process is smooth and stationary (the mean is constant and the covariance function only depends on the distance between the inputs \( x(i) \), that has been proven to work well in practice [16] is:

\[
C(x(p), x(q)) = \nu_1 \exp \left[ -\frac{1}{2} \sum_{d=1}^{D} \omega_d (x^d(p) - x^d(q))^2 \right] + \nu_0 \delta(p, q),
\]

(1)

where \( x^d(p) \) denotes the \( d \)-th component of the \( D \)-dimensional input vector \( x(p) \), \( \Theta = [\nu_1, \omega_1, \ldots, \omega_D, \nu_0]^T \) is the vector of hyperparameters and \( \delta(p, q) \) is the Kronecker operator defined as:

\[
\delta(p, q) = \begin{cases} 1, & p = q \\ 0, & p \neq q \end{cases}
\]

(2)
The $\omega_1, \ldots, \omega_D$ parameters control the scaling of the distances in each input dimension $x^1(i), \ldots, x^D(i)$. The parameter $\nu_1$ is the overall scale of correlations and $\nu_0$ expresses the process noise variance. The exponential term suggests that less distant input vectors lead to highly correlated outputs while more distant inputs generate low correlated outputs. The $\omega_1, \ldots, \omega_D$ parameters can be used to evaluate the relative importance of the corresponding input components (dimensions), i.e., a high or low $\omega_i$ value means that the inputs in dimension $i$ contain high or low information, respectively. There are other forms of covariance functions that can be chosen [1,13], the only restriction being that these covariance functions must generate non-negative definite covariance matrices, for any set of input points.

Assume a statistical model given by:

$$y(k) = f(x(k)) + \epsilon(k),$$

with an additive uncorrelated Gaussian white noise with variance $\nu_0$, $\epsilon \sim \mathcal{N}(0, \nu_0)$. Given a set of training data pairs of input data $X = [x(1), x(2), \ldots, x(n)]$ and the corresponding vector of output data $y = [y(1), y(2), \ldots, y(n)]$ and a Gaussian process prior on $f(x)$, with zero-mean and Gaussian covariance function such as Eq. (1) we wish to get the predictive distribution of $y(n+1)$ corresponding to a new given input $x(n+1)$. For the random variables $y(1), y(2), \ldots, y(n), y(n+1)$ we can write:

$$y, y(n+1) \sim \mathcal{N}(0, K_{n+1}),$$

where $K_{n+1}$ is the covariance matrix made of submatrices as follows:

$$K_{n+1} = \begin{bmatrix} K & k(x(n+1)) \\ k(x(n+1))^T & K(x(n+1)) \end{bmatrix}$$

The matrix $K$ is the $n \times n$ covariance matrix for the training data,

$$K = \Sigma_{pq} = C(x(p), x(q))$$

and the vector

$$k(x(n+1)) = [C(x(1), x(n+1)), \ldots, C(x(n), x(n+1))]$$

is the $n \times 1$ vector of covariances between the training inputs and the new input. The expression

$$K(x(n+1)) = C(x(n+1), x(n+1))$$

is the autocovariance of the new input.

2.1. Predicting with Gaussian processes

The conditional distribution of Eq. (4) allows to obtain the predictive distribution of $y(n+1)$, which is also Gaussian [16]:

$$P(y(n+1) / y, X, x(n+1)) \sim \mathcal{N}(\mu(x(n+1)), \sigma^2(x(n+1)))$$

where $\mu(x(n+1))$ and $\sigma^2(x(n+1))$ are the mean and variance of the Gaussian predictive distribution, and are given by:

$$\mu(x(n+1)) = k(x(n+1))^T K^{-1} y$$

$$\sigma^2(x(n+1)) = k(x(n+1)) - k(x(n+1))^T K^{-1} k(x(n+1))$$

with $K_{n+1}$ being the covariance matrix of the training data.

Fig. 2. Different types of load diagrams (p.u.) for the same day (winter).
\[ \sigma^2 (x(n+1)) = K(x(n+1)) - k(x(n+1))^T K^{-1} k(x(n+1)). \] (11)

We may say that, given a new input vector \( x(n+1) \), the predicted model output \( \hat{y}(n+1) \) is the mean of the Gaussian distribution, i.e., \( \hat{y}(n+1) = \mu(x(n+1)) \) and the uncertainty of this prediction is given by the variance of the Gaussian distribution \( \sigma^2(x(n+1)) \).

The predictive mean Eq. (10) can be interpreted as a weighted sum of the training outputs \( y \), to make a prediction at the test point \( x(n+1) \).

2.2. Training Gaussian processes

To be able to make predictions, based on Eq. (10), the vector of hyperparameters \( \Theta \) has to be provided either as prior knowledge or estimated from the available data. This may be done by maximization of the log-likelihood of the hyperparameters:

\[ \mathcal{L}(\Theta) = \log P(y/X) \]

\[ = -\frac{1}{2} \log(|K|) - \frac{1}{2} y^T K^{-1} y - \frac{n}{2} \log(2\pi). \] (12)

Optimization requires the computation of the derivative of \( \mathcal{L}(\Theta) \) with respect to each parameter \( \omega_1, \ldots, \omega_D, \nu_0, \nu_1 \) of the vector of hyperparameters \( \Theta \), which is given as:

\[ \frac{\partial \mathcal{L}(\Theta)}{\partial \Theta_j} = -\frac{1}{2} tr \left[K^{-1} \frac{\partial K}{\partial \Theta_j}\right] + \frac{1}{2} y^T K^{-1} \frac{\partial K}{\partial \Theta_j} K^{-1} y, \] (13)

where \( tr \) denotes the trace of the matrix. The calculation of the log-likelihood derivatives involves the \( K \) matrix inversion and takes time of the order \( O(N^3) \), which can be extremely demanding for large data sets. However, for these kinds of data sets, sparse training strategies may be employed to significantly reduce the computational cost [5].

2.3. Method summary

Assume that a training data set is known and consists of a \( n \times d \) matrix \( X \) of input measurements and a \( n \times 1 \) vector \( y \) of output or target values. Use the training data set to develop a model that can be used to make prediction with new data. Assume that the new data, called the testing data, is given by an \( 1 \times d \) input vector \( x^* \). The \( y^* \) represents the target value corresponding to \( x^* \). The goal is to predict the value of \( y^* \) given \( X \), \( y \), and \( x^* \).

In the Gaussian process approach the prediction of \( y^* \) involves the selection of a covariance function \( C(x(i), x(j)) \), where \( x(i) \) and \( x(j) \) are vectors with \( d \) components. It is required that the covariance function be positive semidefinite [13]. The covariance function can be used to construct the \( n \times n \) covariance matrix \( K \) with entries \( K_{ij} = C(x(i), x(j)) \), where \( x(i) \) and \( x(j) \) are rows of \( X \), and also the \( 1 \times n \) cross covariance vector \( k \) with entries \( kj = C(x^*, x(j)) \). The prediction \( y^* \) of \( y^* \) is given by the Gaussian processes Eq. (10).

Besides the prediction \( y^* \), the Gaussian process approach also leads to the prediction estimated variance Eq. (11), where \( K(x^*) = C(x^*, x^*) \). Note that the hyperparameters of the covariance function are not known in advance and they can be estimated via the maximization of the log-likelihood Eq. (12) using the training data.

2.4. Dynamic systems

Gaussian estimator processes can be used for the modeling of dynamic systems of the form Eq. (3) if delayed input and output signals are used as regressors [8, 11], such as

\[ x(k) = [y(k-1), y(k-2), \ldots, y(k-L), u(k-1), u(k-2), \ldots, u(k-L)] \]. (14)

In these cases an autoregressive model is considered, such that the current output depends on previous outputs \( y \), as well as on previous exogenous variables \( u \), up to a given lag \( L \).

The Gaussian process model not only describes the dynamic characteristics of the system but also provides information about the confidence in the predictions. This advantage can be used to point out predictions of poor quality, indicated by corresponding variance high values.

For further details, a presentation of Gaussian processes can be found in [13].

3. The choice of input regressor

The structure and the performance of a GP model is determined by the selection of the covariance function and the regressor, i.e., the components of the input vector \( x(t) \). Usually the definition of the input vector structure, involves some options with a certain degree of arbitrariness. Different types of endogenous and exogenous variables are normally applied. Some
authors [2], include temperature or humidity in their models, others use some auxiliary variables like sinusoidal functions in order to include the periodic characteristics of load behaviour [3,6]. Actually, there is no guaranteed rule that one can follow in this process. It mainly depends on experience and is carried out almost entirely by trial and error. In the approach presented here, the composition of the input vector was specified through active power autocorrelation coefficients analysis and the consequent choice of the active power past values \( y(t-1), \ldots, y(t-n) \) [15].

To test the performance and evaluate the influence of a input vector with derivative terms in the forecast quality a comparative study with different compositions of regressors was made.

The first input vector \((R#1)\) includes two derivative terms \((dt_1 \text{ and } dt_2)\) relative to the load variation in homologous periods in the two previous weeks \((y(t-169) \text{ and } y(t-336))\) (Fig. 3). The inclusion of this terms provides information regarding the expected behavior of the load \( y(t) \). The final structure of \(R#1\) is defined as follows

\[
X_1(t) = [y(t-1), y(t-2), dt_1, dt_2],
\]

with \( dt_1 = y(t-336) - y(t-337) \) and \( dt_2 = y(t-168) - y(t-169) \).

A second input vector \((R#2)\), without the derivative terms, was specified. The autocorrelation evolution [15] allows to identify the highest correlation values for the previous two contiguous load values \((y(t-1) \text{ and } y(t-2))\), and for the two homologous values in the two previous weeks \((y(t-168) \text{ and } y(t-336))\). Thus, the second regressor structure is defined as:

\[
X_2(t) = [y(t-1), y(t-2), y(t-168), y(t-336)].
\]

As the two contiguous values around \( p(t-168) \) and \( p(t-336) \) also show significant correlation coefficients the inclusion of the values \( p(t-167), p(t-169), p(t-335) \) and \( p(t-337) \) in the input vector allows to capture information on the consumption trend and is designated by “tendency concept” [15].

The final structure of a third input vector that includes besides the \( R#2 \) terms the “tendency concept” was set as:

\[
X_3(t) = [y(t-1), y(t-2), y(t-167), y(t-168), y(t-169), y(t-335), y(t-336), y(t-337)].
\]

4. Simulation results

The GP model was trained, tested and validated, for all the different load diagrams types presented in the case study. To evaluate and compare the quality of the forecast results, the values obtained for some of the most current statistical indicators were analysed. The mean absolute percentage error (MAPE) value is the most common error indicator, as is generally accepted for comparing different forecast approaches and is defined by

\[
\text{MAPE}(\%) = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{\hat{y}_t} \right| \times 100,
\]

where \( y_t \) is the load value at instant of time \( t \), and \( \hat{y}_t \) is the corresponding forecast value. Other statistical indicators are also necessary to provide a more comprehensive view of forecast results such as the percentage
Fig. 4. Service sector – Forecast results during two weeks in a winter period (regressor R#3).

Fig. 5. Service sector – Forecast results for two different types of day (regressor R#3): a) weekend; b) weekdays.

error \( (PE) \) and the mean percentage error \( (MPE) \), given by

\[
PE_t = \frac{y_t - \hat{y}_t}{\hat{y}_t} \times 100
\]

\[
MPE = \frac{1}{n} \sum_{t=1}^{n} PE_t
\]

where \( n \) is the size of the sampling in the analysis. It is desirable that these statistical indicators should not deviate much from zero, as a sign of lack of bias in the forecast series of values.

Table 1 presents the forecasting MAPE and MPE results for the three regressors used in the experiments for the service sector. According to Table 1 the best mean absolute percentage error (MAPE) obtained dur-
Fig. 6. Service sector – Percentage error PE for the three regressors: a) R#2; b) R#1; c) R#3.

Fig. 7. Service sector – The 95% confidence region for the Forecast values (third regressor).

Table 1
Service sector – Simulation error results

<table>
<thead>
<tr>
<th>Regressor</th>
<th>MAPE (%)</th>
<th>MPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R#1</td>
<td>2.03</td>
<td>-0.055</td>
</tr>
<tr>
<td>R#2</td>
<td>2.57</td>
<td>0.476</td>
</tr>
<tr>
<td>R#3</td>
<td>1.39</td>
<td>0.102</td>
</tr>
</tbody>
</table>

The forecast and the actual values, for this last regressor, are shown in Fig. 4. The details of the first two days (weekend) and last two days (weekdays) of Fig. 4 are presented in Fig. 5. We can also see, from Table 1, that the best value of MPE (0.055%), is obtained with the regressor R#1. However, Figs 6-b) and c) show that this result is a consequence of PE\(_t\) symmetry around zero for regressor R#1 rather than PE\(_t\) low values as it happens with regressor R#3.

Another important statistic indicator is the estimate of the forecast standard deviation (calculated with (11)). Figure 7 represents the actual active power val-
ues and the 95% confidence band (shaded area), obtained with the third regressor, for the forecast values. Almost all the actual values are included in this interval (94% of the actual values). Figure 8 shows the estimate of the forecast standard deviation for all the regressors. The results clearly indicate a higher confidence in the forecasts obtained with regressor R#3 than with the other two regressors.

In what concerns the non-residential and residential sectors the results obtained (see Tables 2 and 3) are very similar but a bit worse than those obtained for the service sector.

As for the service sector regressor R#3 shows better MAPE values than the others regressors in both sectors. Regressor R#2 has a better MPE value for the residential sector because of its PEt symmetry around zero, not from lower level PEt values than the others regressors. All the MAPE results are in line with other approaches [10].

According to several published values of MAPE [10] an error margin of 2% to 5% for a certain period is usually accepted as adequate for STLF predictions. Models based on regressive ANN methodologies, for the same case study presents a major level of error [15]. The results obtained with this type of GP model present more accuracy, for the same load patterns.

Moreover, the ANN methodologies are more complex during the training process. Training, testing and validation of the neural network for this case study takes several hours. The GP model is faster, it takes only a few minutes for the training process.

5. Conclusions

The application of short-term methodologies in a deregulated environment represents a significant role in the decision support of electrical distribution companies. Gaussian process modeling represents a new approach applied to STLF methodology when applied to the electrical distribution sector. This approach, besides imparting forecasts, also provides information about the confidence levels in the forecasts results. One of the most important steps in the application of this methodology is the vector of regressors selection. In this paper we evaluated the forecast quality change
when a derivative term is included in the regressor. The GP model was trained and tested with real-life data values. The results presented confirm the value of the proposed Gaussian process approach in short-term load forecast. Values of MAPE are consistent and always less than 3.2% for different climatic situations and different patterns. The inclusion of the derivative term in the regressor, presents good results, namely the MPE value that is close to zero, sign of lack of bias. However, in what concerns MAPE our experiments show better values for the regressor that includes the “tendency concept” (R#3) a fact that justifies its use as a preferential regressor.

Acknowledgements

The authors gratefully acknowledge the contributions of EDP Distribuição (Portugal) – Direccão do Centro for all the provided data.

This work has been partially supported by FCT under project grant PEst-C/EEI/UI0308/2011.

References