Computers & Industrial Engineering 197 (2024) 110525

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Energy forecast for a cogeneration system using dynamic factor models

Andrés M. Alonso [a], A.E. Sipols [b], M. Teresa Santos-Martín [c,*]

[a] Department of Statistics and Institute Flores de Lemus, Carlos III University, 28903 Madrid, Spain
[b] Department of Applied Mathematics, Materials Science and Engineering and Electronic Technology, Rey Juan Carlos University, Madrid, Spain
[c] Department of Statistics, Institute of Fundamental Physics and Mathematics, University of Salamanca, Salamanca, Spain

ARTICLE INFO

Keywords:
Dynamic factor analysis
Cogeneration forecast
Clustering
Multivariate time series

ABSTRACT

Cogeneration is used in different sectors of industry and it allows that two types of energy to be efficiently obtained from a single source. Accurate predictions are fundamental to optimize energy production, considering the variability that occurs in the daily market. This study adjusts and predicts cogeneration using real data from a Spanish energy technology center, using dynamic factor analysis methodology and incorporating covariates such as temperature and relative humidity. A comparative analysis is performed to evaluate the improvements achieved by implementing cluster-structured dynamic models versus other methods. Furthermore, a robust interpolation method has been implemented to handle missing data in both the main variable and the covariates.

1. Introduction

The simultaneous production of two or more forms of energy, usually electricity and heat, is known as cogeneration (CHP). This type of production takes advantage of the excess heat generated during the process, which would be lost in other electricity generation systems. Therefore, this type of energy production is an efficient and sustainable option to consider.

The fundamental objective of cogeneration is to avoid the loss of energy, (He et al., 2021) that occurs in typical electricity generation processes, where excess heat is not reused but released into the environment. Although it is not possible to transform all the heat generated by the thermodynamics of production, some of it can be used for heating, cooling, or other industrial processes, reducing greenhouse gas emissions and increasing energy efficiency.

Cogeneration plants are adapted to energy demand and operate on a variety of fuel sources, such as natural gas, biomass, and waste heat from industrial processes. The growing demand for sustainable and efficient energy production has led to an increasing interest in cogeneration technologies, which have an important contribution to make to energy production while reducing environmental impact.

Recently, there has been a growing interest in the development of cogeneration forecasting models that utilize both meteorological and energy consumption data to optimize cogeneration production and minimize daily market offer deviations, (Ebrahimi-Moghadam et al., 2021). Various research studies have addressed the development of forecast models for CHP using meteorological and energy consumption data.

The most commonly used techniques for fitting and predicting temporal data can be categorized into two main groups: classical methodology, and machine learning. Currently, it is common to find research studies that compare methods from both categories. Miroshnyk et al. (2021), Weber et al. (2019) use neural networks for renewable energy and electricity forecasting. Chung et al. (2022), Sarathkumar and Goswami (2022) utilize machine learning and parallel CNN-LSTM attention for district heater load forecasting and renewable energy resource forecasting for a virtual power plant in the electricity market. Nepal et al. (2020) and Fan et al. (2019) employ clustering and ARIMA models, as well as various machine learning techniques including hybrid models, for energy management in buildings. Deb et al. (2017) offer a comprehensive review of time series forecasting techniques for building energy consumption. Bedi and Toshniwal (2019) present a deep learning framework for electricity demand forecasting. Klyuev et al. (2022) provide a comprehensive overview of electric energy consumption forecasting methods. Chaturvedi et al. (2022) conduct a benchmarking study of time series models for energy demand forecasting in India, with the Fb Prophet model. Runge and Saloux (2023) compare artificial intelligence models for district heating demand forecasting, specifically using LSTM and XGBoost models. Shaikh et al. (2023) propose a temporal convolutional network for seasonal energy consumption forecasting, addressing the limitations of recurrent neural networks. Ni et al. (2024) develop deep learning-based models for building energy forecasting.

In Deng et al. (2022), gas distribution quality in high-temperature proton exchange membrane fuel cells is examined using data-driven

* Corresponding author.
E-mail addresses: amalonso@est-econ.uc3m.es (A.M. Alonso), anaelizabeth.garcia@urjc.es (A.E. Sipols), maysam@usal.es (M.T. Santos-Martín).

https://doi.org/10.1016/j.cie.2024.110525
Received 11 February 2024; Received in revised form 21 June 2024; Accepted 24 August 2024
Available online 28 August 2024

surrogate models. The framework proposed by Teichgraeber and Brandt (2019) utilizes clustering methods to model time-varying operations in complex energy systems optimization problems. Mezzi et al. (2021) propose an approach using Echo State Network for the prognostics of proton exchange membrane fuel cells under variable load conditions. Finally, Ikeda and Nagai (2021) propose a hybrid algorithm combining metaheuristics and machine learning to optimize daily operating schedules in building energy systems.

In other fields, these methodologies are currently widely used. For example, Lee and Kang (2024) propose a dynamic method that combines predictions from locally trained neural networks in a decentralized environment. He et al. (2023) use LSTM for forecasting flight reservation demand. Yan et al. (2024) focus on public health emergencies employing a Spatiotemporal Multigraph Convolutional Network (SMEGCN). Taşcı et al. (2023) predict the lifespan of production equipment using machine learning and hybrid models.

To achieve more accurate predictions of cogeneration data, it is proposed to use Dynamic Factor Analysis (DFA), which allows modeling the relationship between observed variables and latent factors that change over time. This approach enables the examination of how unobserved factors affect measured variables at different temporal points. Alonso et al. (2016) employ Dynamic Factor Models on multivariate time series, applied to European industrial production indices. García-Martos et al. (2012) use DFA to forecast electricity prices, taking into account the multivariate structure of the data. Dordonnat et al. (2012) discussed a dynamic periodic multivariate regression model for hourly electricity data.

Dynamic Factor Models with Cluster Structure (DFMCS) assign different latent factors to distinct groups of variables, thereby capturing heterogeneity and providing a more detailed and specific analysis for each cluster. In this context, Alonso et al. (2020) presented a procedure to fit (DFMCS) to heterogeneous time series data that may include multivariate additive outliers and level shifts for electricity market. Vialetto and Noro (2020) presented an innovative approach based on big data analysis and cluster analysis to design cogeneration systems that can suit energy demand profiles more efficiently, choosing the correct type of cogeneration technology, operation strategy and, if necessary, the size of energy storage. A case study based on a cogeneration plant is analyzed, showing that the proposed method is useful for designing cogeneration systems for industry and allowing for energy and economic savings. Bujalski and Madejski (2021) introduced a new methodology using a big data-driven model for short-term forecasting of heat production in combined heat and power plants. The methodology accurately predicts hourly heat load in the day-ahead horizon, allowing for better planning and optimization of energy and heat production by cogeneration units. Ifaei et al. (2023) provided a comprehensive review of the major applications and remaining challenges of machine learning in sustainable energies, focusing on prediction, clustering, and optimization, as well as multi-carrier energy systems, spatial–temporal analytics, and circular integration.

In the present study, the objective is to make cogeneration predictions using the data provided by a technology energy centre for a specific area. The aim is to minimize the deviations in offers in the market by implementing a dynamic factor analysis approach. The advantage of using dynamic factor analysis for predicting cogeneration data is that it allows for the identification of underlying factors that contribute to the observed variability in the data. Dynamic Factor Analysis (DFA) has been used to model the time-varying relationship between meteorological variables (such as temperature and humidity) and CHP production. In addition, it is complemented by the Dynamic Factor Model with Cluster Structure (DMFCS), which takes into account features like heterogeneity and cluster structure, improving the analysis of CHP data and providing valuable information on the underlying factors and their impact on the system.

Through a case study of the electricity market, the effectiveness of the approach is demonstrated in terms of factor and loading estimation, outlier cleaning, and the utilization of cluster structure for understanding and forecasting.

Fig. 1 presents a schematic of the research framework. One of the challenges encountered when working with real-world data is the presence of missing data points, which can lead to inaccuracies in predictions. To address this, the local median interpolation method is employed, providing robust estimates and reducing sensitivity to outliers. In general, capturing the temporal pattern and seasonal variations of the series requires a prior study of the data, as well as the possible factors that may influence such variations. Using Dynamic Factor Analysis (DFA) integrated with seasonal ARIMA models allows us to effectively capture both temporal dynamics and seasonality. Identifying and exploiting cross-correlation structures between multiple time series is essential to obtain better forecasts. In this case, Dynamic Factor Models with Cluster Structure (DFMCS) are designed to capture and utilise cross-correlation structures, allowing models to take into account inter-dependencies between different time series, improving forecasting accuracy. There are several time series adjustment methodologies available to predict future data, but it is common for the long-run predictions of these models to be inaccurate. For this reason, the paper tests different fitting models that integrate dynamic latent factors with external covariates. The resulting models effectively capture both short-term fluctuations and long-term trends. This dual capability allows for more reliable long-term forecasts, mitigating the impact of uncertainties.

In summary, the main contributions of this paper are the introduction of an innovative modeling strategy based on DFA and DFMCS for forecasting CHP production, with significant improvements in prediction accuracy compared to competing models. The use of seasonal ARIMA models in DFA to capture the temporal dynamics of common factors and cross-correlation structures between time series, as well as the implementation of the local median interpolation method to handle missing values, improving the accuracy of the data. The proposed models have the potential to contribute to the advancement of cogeneration as a viable alternative in energy production and to promote the use of innovative techniques for data analysis in this field. This is a significant advantage, as it allows for better planning and decision-making in the energy production process.

The rest of the paper is organized as follows: Section 2 describes the statistical methodology. The results of the statistical methods applied to the energy data are discussed in Section 3. Finally, Section 4 presents the conclusions.

2. Methodology applied to the forecast and estimation model

In this section, the main statistical methods used to adjust and predict the energy data generated by an electrical plant are presented following the scheme depicted in Fig. 1. Firstly, a robust interpolation method for missing data is introduced for all variables involved in the models. Subsequently, a detailed explanation of the Vector Autoregression, Dynamic Factor Model, and Dynamic Factor Models with Cluster Structure used in the study is provided, concluding with the commonly used metric to evaluate the accuracy of the models.

2.1. Interpolation

A first difficulty when performing data analysis is the presence of missing values both in the variable to be predicted and in the possible covariates that will be used in the predictive model. There are several interpolation methods for missing data in time series, including the most common ones such as linear interpolation, moving average, polynomial interpolation, spline interpolation, and the geometric mean method. The choice of method depends on the nature of the data and the objective of the analysis, and it is important to note that

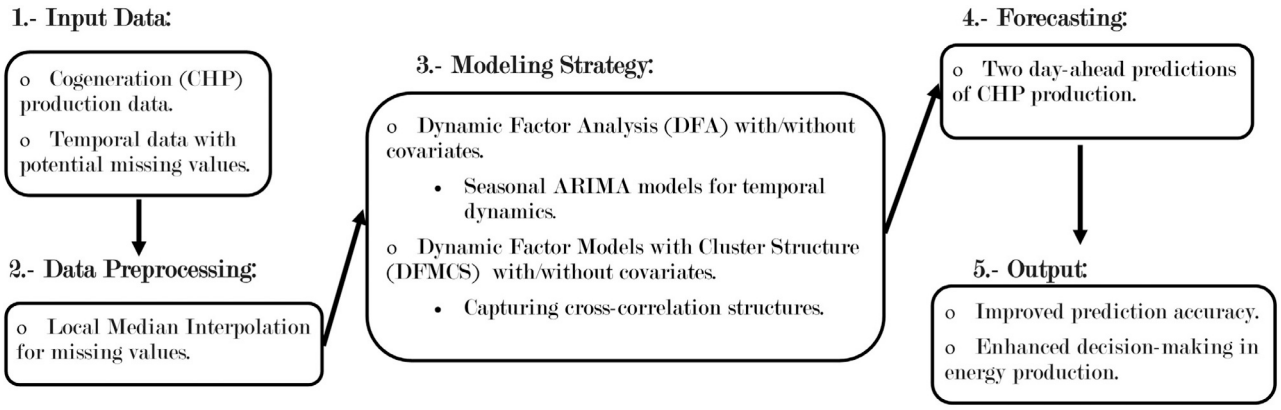

Fig. 1. Scheme of the research framework.

interpolation methods may not be appropriate in all cases, especially if the time series has a complex trend or seasonal pattern, (Box et al., 2015; Chatfield, 2013).

In this paper, a robust missing data imputation procedure has been implemented that takes into account the dependency structures of the variables. The interpolation method uses the local median. The first step of the procedure is to consider the daily time series, $\mathbf{y}_t = (y_{t,1}, y_{t,2}, \dots, y_{t,24})'$, that is, $\mathbf{y}_t$ contains the 24 values of the hours of day t. If there is any missing value in $\mathbf{y}_t$ then a window of d days and h hours is taken around the missing observation. Suppose the value $y_{t,i}$ is missing, the following set $\{y_{t-d,i-h}, \dots, y_{t-d,i+h}, \dots, y_{t+d,i-h}, \dots, y_{t+d,i+h}\}$ is considered, and the median of these values is taken. In the few cases in which all values from the previous set were missing, the size of the days window was increased, keeping the number of hours in the corresponding window constant. Taking small windows in both days and hours ensures that both hourly and daily dependence are taken into account. Using the median instead of the mean protects us from the presence of outliers. The method is applied in a loop over all missing values in the dataset. The result is a dataset in which the missing values are replaced by local medians. Of course, other interpolation procedures can be used that, for example, include the dependence between the three time series, but the small number of missing values did not make this necessary.

2.2. Vector Autoregression

The Vector AutoRegression (VAR) model is a time series analysis technique that enables the modeling of the relationship between multiple variables over time. The model assumes that each variable depends on its past values and the past values of other variables. In this way, the interdependence dynamics between the variables can be captured and used to make future predictions.

VAR models have been extensively used in various fields, such as economics, finance, and engineering, due to their ability to capture complex relationships between multiple variables over time. For instance, VAR models have been applied to study the causal relationships between economic variables (Enders, 2004; Lütkepohl, 2009), forecast exchange rates, and analyze the dynamics of water quality parameters in a river system.

In the described case, the VAR model is utilized to predict energy generation based on temperature and humidity. The inclusion of these covariates in the model allows for a more accurate prediction of energy generation, as they are known to influence energy production. The use of VAR models in energy forecasting has been widely researched in recent years, with studies focusing on applications such as wind power, solar power, and energy demand. The general formulation for a VAR(p) model is:

$$\mathbf{y}_t = \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \mathbf{\Phi}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\mathbf{y}_t$ is a m-dimensional time series, $\mathbf{\Phi}_i$ are $m \times m$ matrices and $\boldsymbol{\varepsilon}_t$ is a m-dimensional white noise process. It should be noticed that the estimation of matrices $\mathbf{\Phi}_i$ involves pm^2 parameters. The VAR(p) model is stable or stationary if all zeroes of the determinant equation $|\mathbf{I} - \sum_{i=1}^{p} \mathbf{\Phi}_i B^i| = 0$, where $\mathbf{I}$ is the $m \times m$ identity matrix, are greater than 1 in absolute value (see, for instance, Lütkepohl, 2005).

Overall, the VAR model is a powerful tool for modeling the interdependence between multiple variables over time and making predictions for future values. In Section 3, it is assumed that the vector $\mathbf{y}_t$ is a 3-dimensional time series conformed by three univariate time series: generated energy, temperature and relative humidity.

2.3. Dynamic Factor Model

Dynamic Factor Model (DFM) is a statistical technique used to identify and model time-varying patterns in a multivariate time series of data (see, for instance, Peña & Box, 1987 and Stock & Watson, 1988). In DFM, the time series data is decomposed into a linear combination of a set of unobserved factors that are assumed to be responsible for the observed variation in the data. These factors are modelled as stochastic processes that evolve over time according to a set of dynamic equations. The model parameters can be estimated using maximum likelihood estimation or by principal components (see, Stock & Watson, 1988), and the resulting model can be used to forecast future values of the time series. DFA has been applied to a wide range of fields, including finance, economics, engineering, and environmental science, among others. The DFM can be expressed as

$$\mathbf{y}_t = \mathbf{\Lambda} \mathbf{f}_t + \boldsymbol{\varepsilon}_t, \tag{2}$$

where $\mathbf{y}_t$ is a m-dimensional time series, $\mathbf{f}_t$ is a r-dimensional time series called *common factors*, $\mathbf{\Lambda}$ is the loading matrix with dimension $m \times r$ that relates the set of r common unobserved factors with the vector of observed series $\mathbf{y}_t$, and $\boldsymbol{\varepsilon}_t$ is an m-dimensional vector of innovations also called *specific factors*. Also, in this work the innovation are assumed to be white noise and no model is employed to fit them. For instance, a vector $\boldsymbol{\varepsilon}_t$ could be normal with zero mean and diagonal variance–covariance matrix $\mathbf{S} = E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t')$. Alternatively, when employing Dynamic Factor Analysis (DFA) the innovation can be modelled as independent auto-regressive processes.

It is important to note that, in Section 3, the DFM model is applied to the daily multivariate time series that results from considering the 24 time series formed by the hourly measurements. That is, from a univariate time series

$$\mathbf{Y} = \{y_1, \dots, y_{24}, y_{25}, \dots, y_{48}, \dots\}, \tag{3}$$

a multivariate time series is obtained

$$\widetilde{\mathbf{Y}} = \{\mathbf{y}_1, \mathbf{y}_2, \dots\}, \tag{4}$$

where $\mathbf{y}_1 = (y_1, \ldots, y_{24})'$, $\mathbf{y}_2 = (y_{25}, \ldots, y_{48})'$ and so on (see, for instance, Ramanathan et al., 1997). The ith time series of the vector, $\mathbf{y}_t$, are the measurements at hour i of each day. Therefore, the vector $\mathbf{y}_t$ has dimension 24×1. The procedure defined by expressions (3) and (4) are also called parallel approach. It should be noted that the parallel approach is recommended when the way in which the univariate time series is generated has a block generation component, as is this case. The production company can decide in advance what its production schedule (P_S) will be for the next 24 h. That is, the 24 values $P_{S,1}, P_{S,2}, \ldots, P_{S,24}$ are decided simultaneously. The final production (P) of the next day depends on the programmed values and the weather conditions, that is, P (the observed univariate time series) is a function of P_S (a vector time series) and other exogenous variables. The use of a VAR(p) model for a vector of this dimension is not recommended because it requires estimating $p + 1$ matrices of dimension 24×24, the p autoregressive matrices and the covariance matrix of the noise process.

The common unobserved factors, f_t, can be non-stationary, including regular or seasonal unit roots and also auto-regressive and/or moving average regular and seasonal components. Here, this approach is followed, and, thus, factors are modelled to follow seasonal ARIMA $(p, d, q) \times (P, D, Q)s$ process which are used to obtain the factors forecasts, and from them the energy forecasts. For instance, the ith factor $f_{t,i}$ would be modelled by

$$(1 - B)^d (1 - B^s)^D \phi_i(B) \Phi_i(B^s) f_{it} = \\ = c_i + \theta_i(B) \Theta_i(B^s) w_{it}, \tag{5}$$

where $\phi_i(B) = (1 - \phi_{i1} B - \phi_{i2} B^2 - \cdots - \phi_{ip_i} B^{p_i})$, $\Phi_i(B^s) = (1 - \Phi_{i1} B^s - \Phi_{i2} B^{2s} - \cdots - \Phi_{iP_i} B^{P_i s})$ are the regular and seasonal stationary autoregressive polynomials, $\theta_i(B) = (1 - \theta_{i1} B - \theta_{i2} B^2 - \cdots - \theta_{iq_i} B^{q_i})$ and $\Theta_i(B^s) = (1 - \Theta_{i1} B - \Theta_{i2} B^{2s} - \cdots - \Theta_{iQ_i} B^{Q_i s})$ are the invertible regular and seasonal moving averages polynomials, and B is the lag operator such that $B y_t = y_{t-1}$.

The roots of $|\phi_i(B)| = 0$, $|\Phi_i(B^s)| = 0$, $|\theta_i(B)| = 0$, $|\Theta_i(B^s)| = 0$, satisfy the usual stationarity and invertibility conditions, and w_{it} are identically distributed and uncorrelated random variables, that is $E(w_{it} w_{it-h}) = 0, \forall h \neq 0$. It is also assumed that the error term of the common factors w_{it} is uncorrelated with the specific factors, that is $E(w_{it} \epsilon'_{t-h}) = 0, \forall h$. The term c_i is the intercept of the model for the common factors, and its inclusion in (5) can be particularly relevant to calculate long term forecasts in the non-stationary case. When exogenous variables are integrated into the ARIMA model to explain the behavior of the time series, it is referred to as ARIMAX models.

A two-stage procedure is used to estimate the DFM model defined by (2) and (5). The factors and the loading matrix in (2) are estimated by a principal components procedure and the coefficients of ϕ_i, Φ_i, θ_i and Θ_i in (5) for $i = 1, 2, \ldots, r$ are estimated by maximum likelihood procedure. Once these elements have been estimated, they can be used to make predictions of future values for the factors. After that, the prediction for the multivariate time series $\mathbf{y}_t$ is obtained using relation (2).

2.4. Dynamic factor models with cluster structure

Consider a vector of zero-mean stationary time series, denoted as $\mathbf{y}_t = (y_{1t}, \ldots, y_{mt})'$. It is assumed that each element of the observed series vector is a linear combination of global and specific components within k clusters or groups, with the presence of some noise. These factors can be represented as follows:

- The global factors are denoted by the r_0-dimensional vector $\mathbf{f}_{0t} = (f_{01t}, \ldots, f_{0r_0t})'$.
- The global factor loading matrix is represented by $\mathbf{\Lambda}_0 = [\mathbf{\Lambda}'_{0,1} \mid \cdots \mid \mathbf{\Lambda}'_{0,s}]'$, with dimensions $k \times r_0$. Here, $\mathbf{\Lambda}_{0,i}$, for $i = 1, \ldots, s$, corresponds to the $k_i \times r_0$ loading matrix for the k_i series of the ith group.
- The specific factors for the ith cluster are expressed as the r_i-dimensional vector $\mathbf{f}_{it} = (f_{i1t}, \ldots, f_{ir_it})'$.
- The matrix of specific factor loadings, which exclusively affect the k_i time series in the ith group, is denoted by $\mathbf{\Lambda}_i = [\mathbf{0}'_{i,1} \mid \ldots \mid \mathbf{\Lambda}'_{i,i} \mid \ldots \mid \mathbf{0}'_{i,s}]$, with dimensions $k \times r_i$.

The (DFMCS) Dynamic Factor Model with Cluster Structure can be expressed as follows:

$$\mathbf{y}_t = \mathbf{\Lambda}_0 \mathbf{f}_{0t} + \sum_{i=1}^{k} \mathbf{\Lambda}_i \mathbf{f}_{it} + \boldsymbol{\epsilon}_t. \tag{6}$$

It should be noticed that the previous model generalizes the DFM since the component $\mathbf{\Lambda}_0 \mathbf{f}_{0t} + \boldsymbol{\epsilon}_t$ is the same as in expression (2) and the component $\sum_{i=1}^{k} \mathbf{\Lambda}_i \mathbf{f}_{it}$ adds the specific dynamics of the series in the clusters.

The fitting procedure of the Dynamic Factor Model with Cluster Structure (DFMCS) involves several steps. Firstly, the observed time series data is cleaned by removing additive outliers, level changes, and outlying time series. Then, the factor loading matrix is estimated by selecting the eigenvectors associated with the largest eigenvalues of the time series' covariance matrix. The factors and common components are computed based on the estimated loading matrix. Next, a clustering algorithm is applied to group the time series according to their linear dependence using the common components. Within each group, new factors and their loadings are estimated using a similar procedure as before. The factors are classified as global or specific based on empirical canonical correlation analysis. The residuals are computed, and the final estimation of factors, groups, and loadings is performed. The procedure aims to minimize the squared error by considering the estimated factors, groups, and loadings. For a detailed methodology, see Alonso et al. (2020) and Ando and Bai (2017).

2.5. Metrics used to evaluate model performance

Mean Absolute Percentage Error (MAPE), is a commonly used metric for evaluating the accuracy of a forecasting model. It measures the average percentage difference between the actual values and the predicted values

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{A_i - P_i}{A_i} \right|, \tag{7}$$

where n is the number of observations to be predicted, A_i is the actual value of observation i, and P_i is the predicted value of observation i. MAPE provides a useful measure of the accuracy of a forecasting model, particularly when the data has a wide range of values. The lower the MAPE, the more accurate the forecasting model.

3. Results of application on energy data

Using hourly energy generation data obtained from a production center in Spain, along with temperature and humidity data collected from a close city between 01/01/2012 and 14/04/2013, the aim is to develop a model for predicting energy generation up to 48 h in advance. The performance of this model will be compared against other models that include temperature and humidity covariates information. The calculations have been programmed using the Matlab program.

The data provided by the technology center consists of 11 280 hourly measurements (470 days × 24 h) of the generated energy, temperature and relative humidity. The period from 01/01/2012 to 30/03/2012 (90 days) is taken as the initial training period, and the rest of the observations, from 31/03/2012 to 14/04/2013, as the testing period of the prediction procedures.

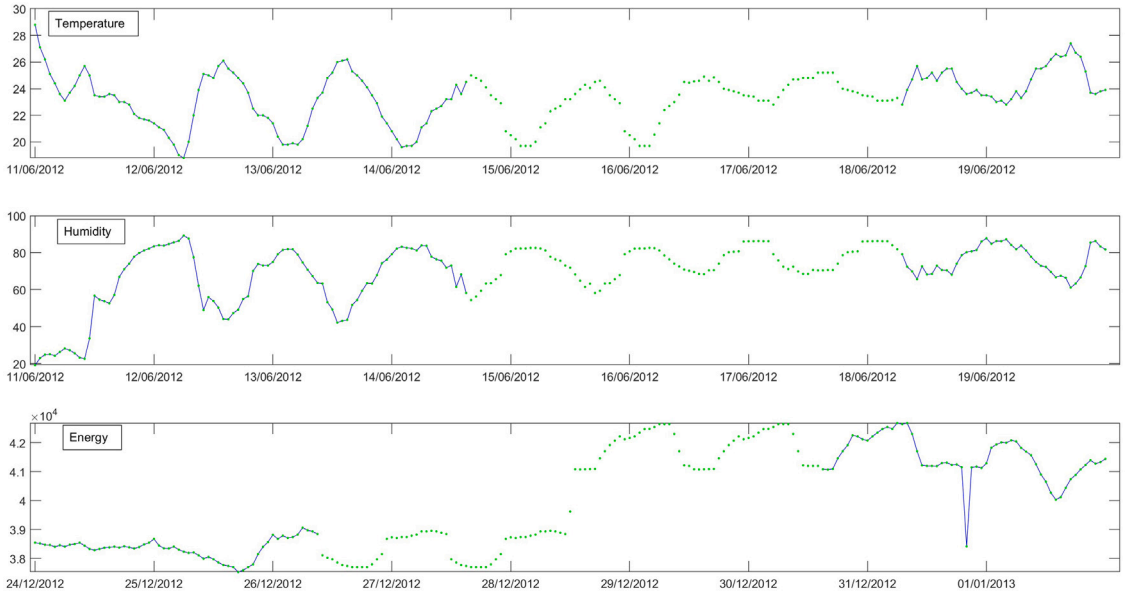

Fig. 2. Examples of missing observation interpolation. Solid line corresponds to real time series and dotted line corresponds to interpolated time series.

3.1. Interpolation

In the data provided by the technology centre, there are 269 missing data for energy and 409 and 407 missing data for temperature and humidity, respectively. As this is a moderate amount of data, it is decided to use a robust interpolation method, which has been described in the methodology section. Fig. 2 shows examples of the interpolated time series for temperature, humidity and energy. It is clear that the local interpolation procedure captures the dynamics of the series, accommodating to possible changes and being robust to the presence of outliers, such as the peak on 31/12/2012.

Table 1
Summary statistics of daily mean absolute percentage errors (MAPE) for temperature and humidity forecasts.

	Temperature		Humidity	
	Day 1	Day 2	Day 1	Day 2
Minimum	1.23%	1.41%	3.58%	3.19%
1st. Quartile	3.26%	3.63%	8.74%	10.42%
Median	6.26%	7.46%	15.72%	17.36%
Mean	8.34%	10.11%	23.80%	26.96%
3rd Quartile	10.60%	13.23%	28.42%	28.59%
Maximum	43.08%	63.25%	222.65%	257.75%

3.2. Exploratory data analysis

Firstly, a study of the correlation structure of the time series of energy production, temperature, and relative humidity has been carried out. Figs. 3 and 4 show the sample autocorrelation and the sample partial autocorrelation of the three time series, respectively. It is observed that the three series have a high regular and seasonal dependence (24 h).

Secondly, the cross-correlation between the energy series and the two series of meteorological variables has been studied. Fig. 5 shows the cross-correlation function using the levels of the series, and Fig. 6 shows this function using the seasonal differences of the series. In this way, the interpretation of the cross-correlation is avoided being biased by the high correlations at lags multiples of 24 as observed in Figs. 3 and 4.

In both Figs. 5 and 6 it is observed that the energy generated has cross dependence on both seasonal (multiples of 24) and regular delays. This is called regency effect and means that energy production depends not only on the temperature at the current time but also on temperatures from previous hours. This effect has been studied by Wang et al. (2016) in the case of electricity demand.

Furthermore, it is interesting to study the cross-correlation structure throughout the year. In Fig. 7, the correlation matrices of the vector $\left(E_{d-1}, E_{d-2}, \ldots, E_{d-24}, T_{d-1}, T_{d-2}, \ldots, T_{d-24}, H_{d-1}, H_{d-2}, \ \ldots, H_{d-24}\right)$ in absolute value, that is, of the 24 daily (d) observations of energy (E), temperature (T) and humidity (H), respectively, are represented. In Fig. 7 the matrices calculated with blocks of three months that correspond approximately to the seasons of winter, spring, summer and autumn are represented. It is observed that the relationship between energy and temperature changes magnitude in the different seasons of the year. The relationship between energy and humidity also presents changes but much smaller. These results justify the choice of the 90-day training period used in Section 3.4.

Finally, the dependence structure of the energy time series for each of the hours of the day has been explored. That is, all observations corresponding to the hth hour of the day are taken to form the hth time series. Fig. 8 shows the autocorrelation functions of these 24 time series, where it can be observed that the series share a certain similarity in their dependence but certain groups of series are visible. This cluster structure will be exploited in the models in the next section.

3.3. Forecasting models

Before adjusting a model to predict energy production, the modeling of the covariates, temperature and humidity, is considered. This will allow us to have predictions of the covariates and obtain energy forecasts in a realistic context. Of course, if the exact location information of the production center could be used, numerical weather prediction models could be employed (see, for instance, Pu & Kalnay, 2019).

1. **Models for temperature and relative humidity prediction**
   Given the high correlation structure between temperature (slightly lower in the case of humidity) and the generated energy, as illustrated in Figs. 5 and 7, modeling of temperature and humidity was considered using time series models that account for the strong hourly seasonality of these variables, as shown in Figs. 3 and 4.

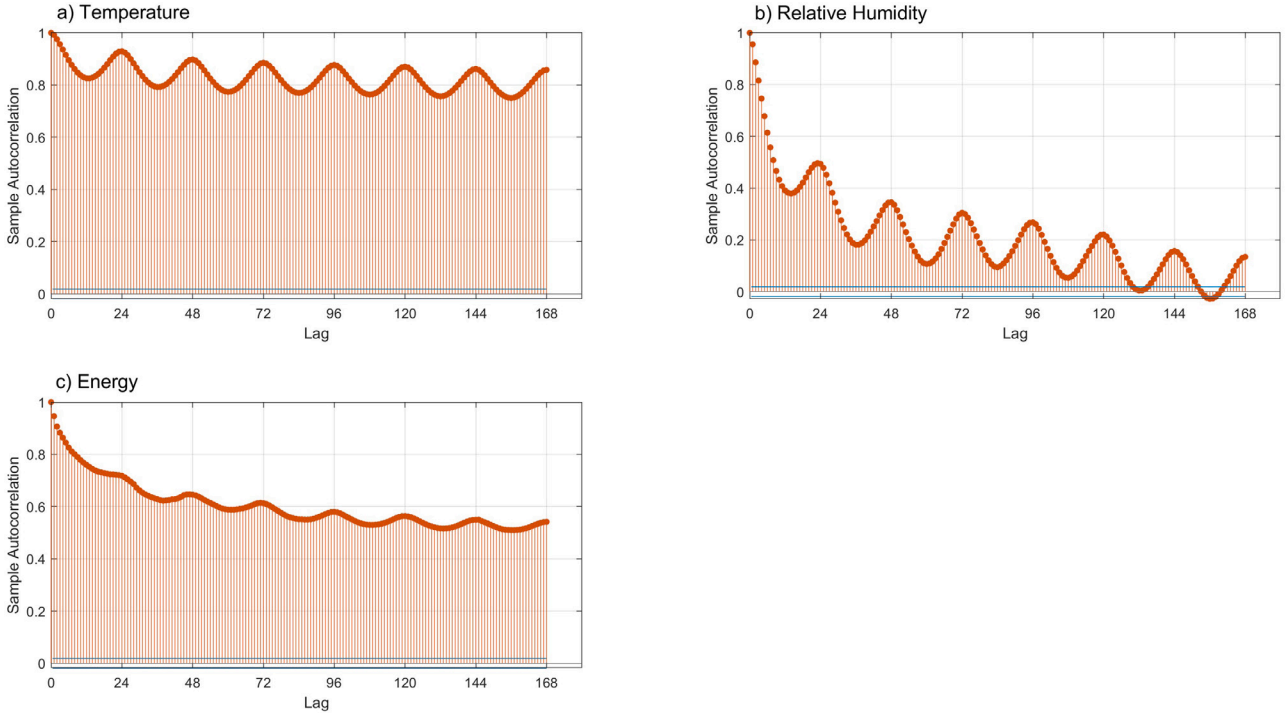

Fig. 3. Autocorrelation functions for (a) Temperature, (b) Relative humidity, and (c) Energy.

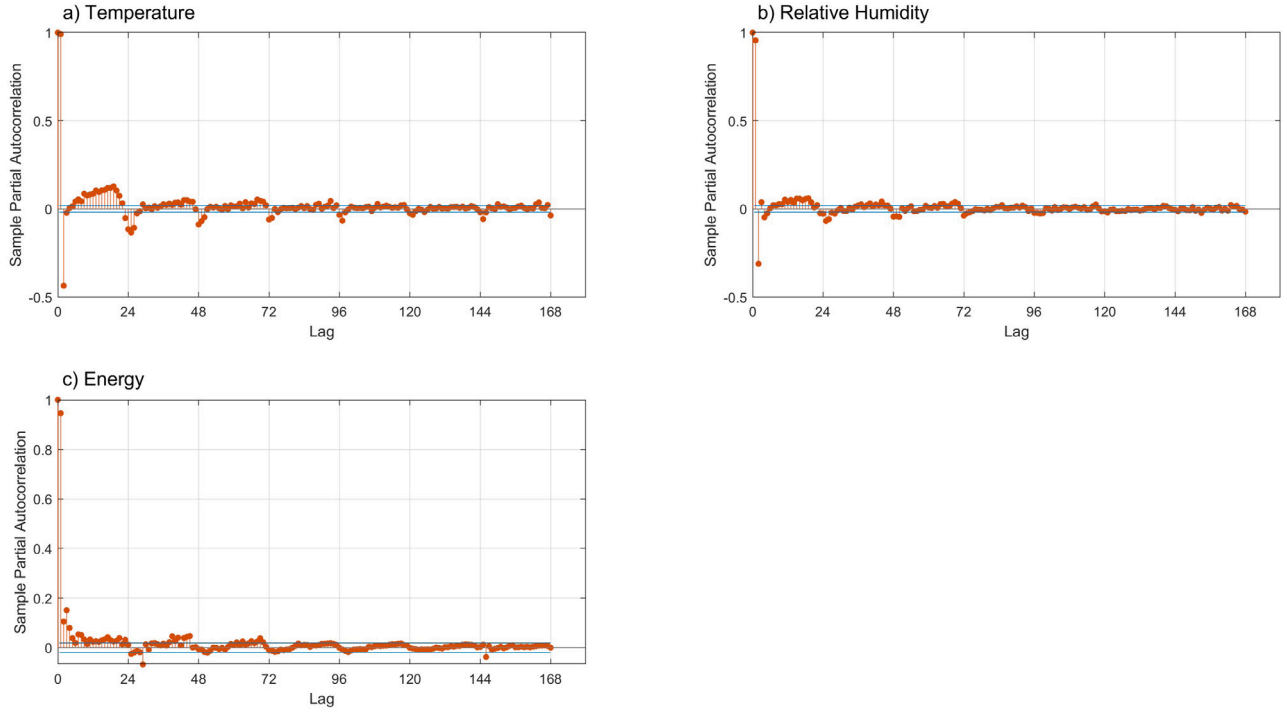

Fig. 4. Partial autocorrelation functions for (a) Temperature, (b) Relative humidity, and (c) Energy.

First, the two vector time series are obtained from the univariate time series (temperature and humidity) using expressions (3) and (4). Thus, two 24×1 dimensional vector time series are obtained. A factorial model will be used to predict each vector time series using a 90-days history and a two-days prediction horizon. If the correlation matrices of the vector $(T_{d-1}, T_{d-2}, \ldots, T_{d-24})$ and the vector $(H_{d-1}, H_{d-2}, \ldots, H_{d-24})$ in Fig. 7 are analyzed, it is verified that the cross-dependence in both temperature and humidity also depends on the season of the year. This justifies the use of a short training period in the considered models. The procedure employs principal component analysis (PCA) to identify underlying factors that explain the variability of temperature and humidity data. Two factors, in the case of temperature, explain around 91.87% of the variability of that series. In the case of humidity, three factors explain around 87.85%. Note that it is advisable not to incorporate 100% of the variability

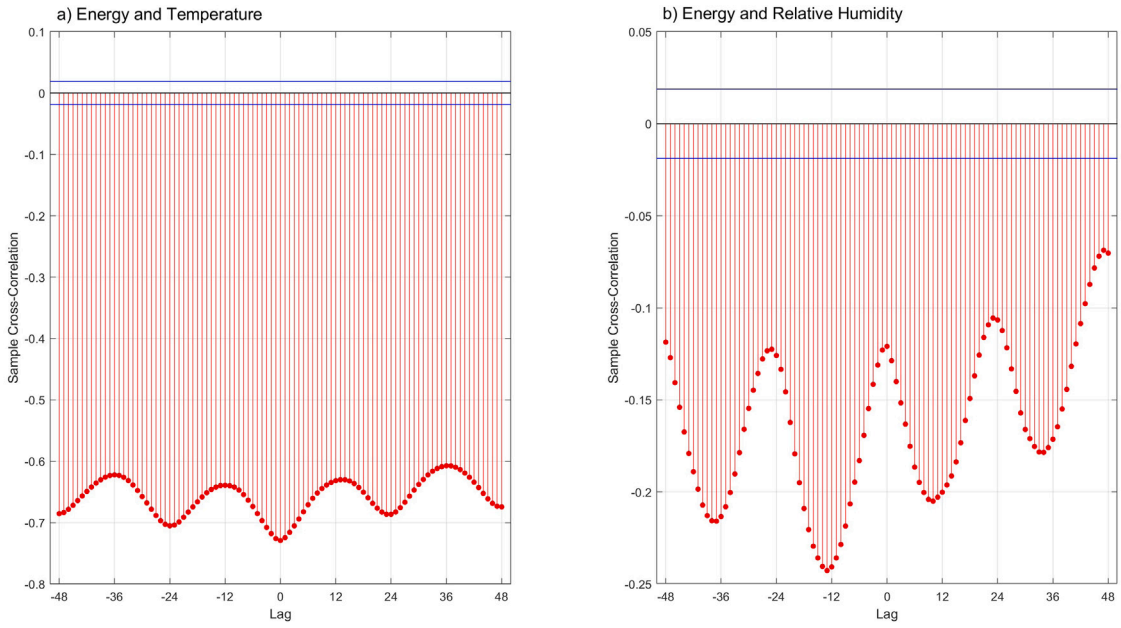

Fig. 5. Cross-autocorrelation functions between (a) Energy and Temperature and (b) Energy and Humidity.

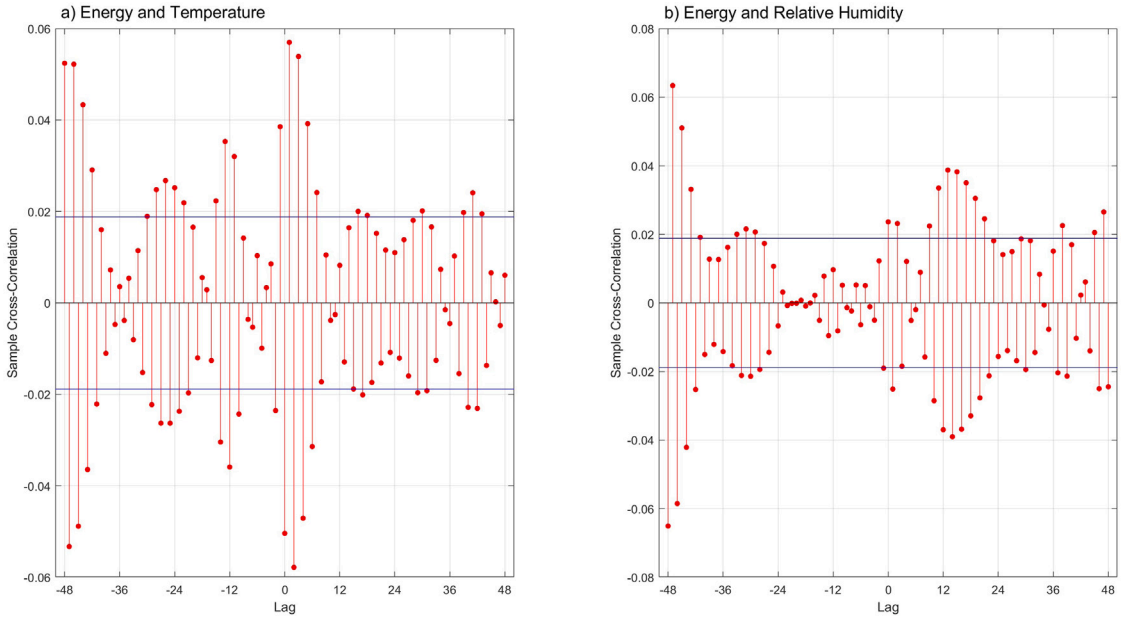

Fig. 6. Cross-autocorrelation functions between the seasonal differentiated (a) Energy and Temperature and (b) Energy and Humidity.

since it introduces random fluctuations that do not improve the predictions. In a sense, extracting the main factors can be understood as a smoothing procedure. The seasonal ARIMA models are selected by the TRAMO-SEATS program (Gómez & Maravall Herrero, 1998) for each factor. Five models are selected and estimated in each training set, one model for each factor. TRAMO-SEATS is a well established procedure included in econometrics software such as EViews and there are interfaces for Matlab and R. It implements the order selection of the ARIMA models, as well as the outliers detection and correction. In Maravall et al. (2015), the authors show that this approach performs well in the context of automatic model identification for a large number of time series particularly in the selection of regular and seasonal difference order. These models enable predictions with a horizon of up to 48 h for these two variables. A summary of the mean daily relative absolute errors for both variables for one- and two-days horizons are shown in Table 1.

2. **Models for the prediction of generated energy**

The following models will be considered to predict the generated energy and their prediction performance will be compared in the next section:

Model 1: A Dynamic Factor Model (DFA) without covariates as the one defined by (2) and (5). Initially, the DFA model is estimated using the generated energy data series, without

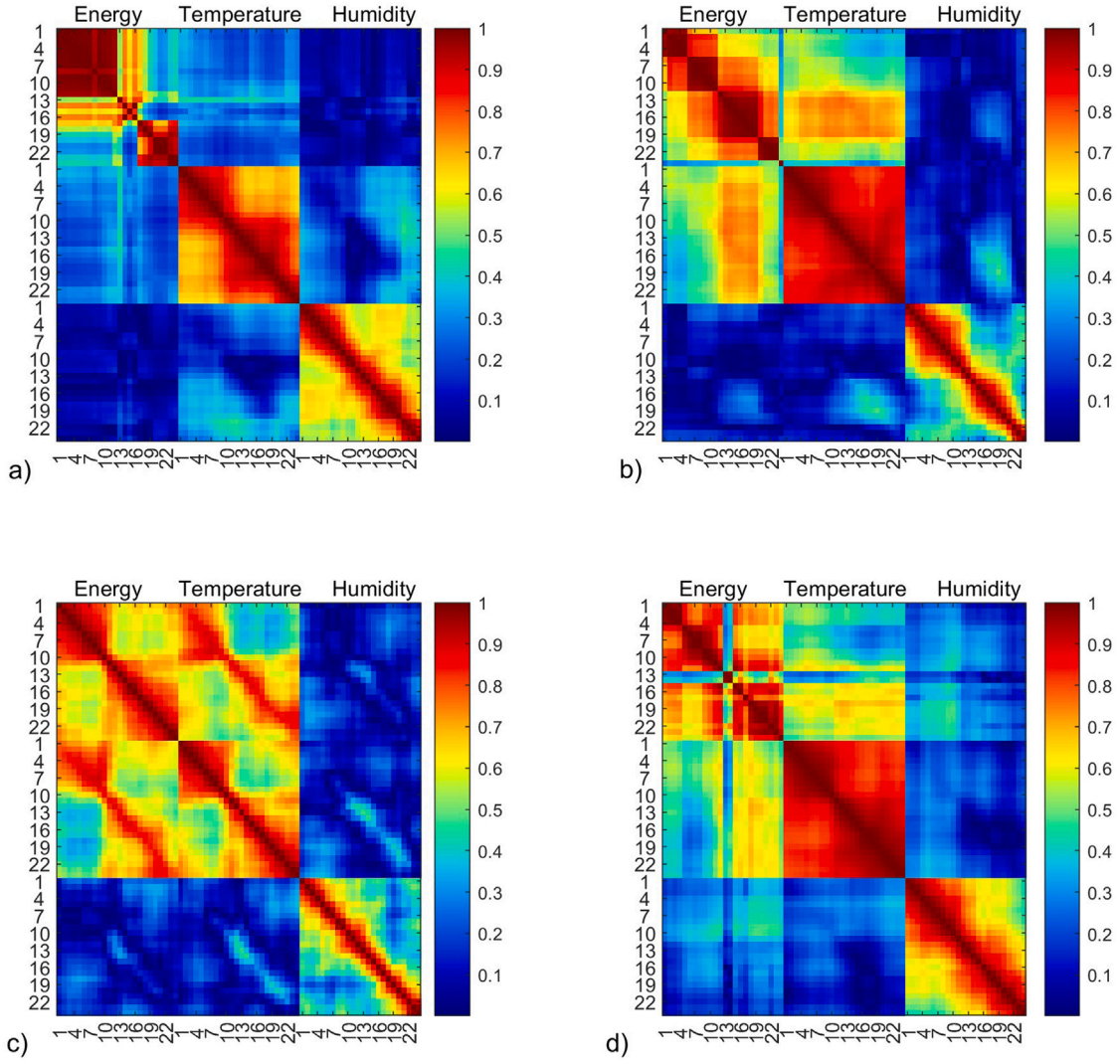

Fig. 7. Correlation (absolute value) matrices in (a) January–March, (b) April–June, (c) July–September, and (d) October–December.

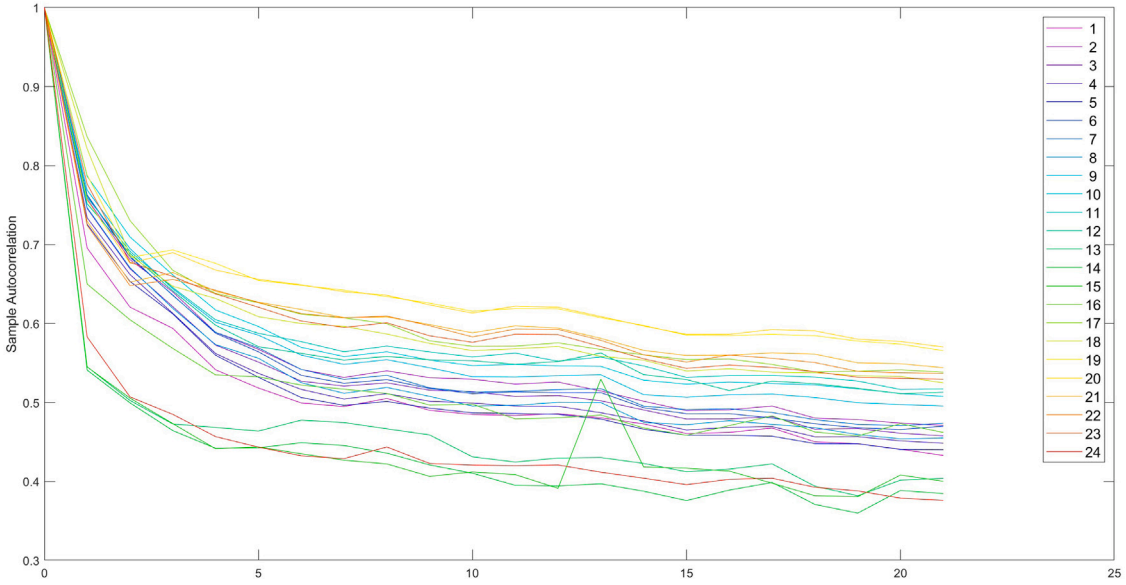

Fig. 8. Autocorrelation functions for the energy time series for each hour of the day.

incorporating the humidity and temperature data. The model has two factors that explain around 98.24% of the variability.

- The dimensions of the vector time series (generated energy) and the factors are 24×1, and 2×1, respectively. That is, $m = 24$ and $r = 2$.

Model 2: A Dynamic Factor Model in which two temperature factors and three humidity factors are introduced as covariates. That is, five covariates are introduced in the seasonal ARIMA model defined by (5). In this case, the predictions of the covariates that have been previously obtained are used.

- As in the previous model, $m = 24$ and $r = 2$. Notice that, in this case, two DFM models are also estimated in order to obtain the predictions of the covariates (temperature and relative humidity). These models have $m = 24$ and $r = 2$ and 3, respectively.

Model 3: This model has the same structure as Model 2, but it uses actual values of the humidity and temperature covariates.

- As in Model 1, $m = 24$ and $r = 2$. Notice that in this case, the covariates are not predicted. They are assumed to be known or reliable external predictions are available.

Remark 1. The fundamental difference between Model 1 and Models 2–3 is that the latter incorporate the information from the covariates. Taking into account the dependence between the energy generated and these covariates (see Figs. 5 and 6) it is expected that Models 2 and 3 outperform Model 1.

Model 1C: A Dynamic Factor Models with Cluster Structure (DFMCS) defined by Eq. (6) without covariates. This DFMCS have the same number of global factors as Model 1 and two specific factors.

- The dimensions of the vector time series (generated energy) and the common factors are 24×1, and 2×1, respectively. There are two clusters that each have one specific factor and a cluster that does not have a specific factor. That is, $m = 24$, $r_0 = 2$, $k = 3$, $r_1 = 1$, $r_2 = 0$ and $r_3 = 1$.

Model 2C: A Dynamic Factor Model with Cluster Structure in which two temperature factors and three humidity factors are introduced as covariates. In this case, the predictions of the covariates that have been previously obtained are used.

- As in the previous model, $m = 24$, $r_0 = 2$, $k = 3$, $r_1 = 1$, $r_2 = 0$ and $r_3 = 1$. In this case, two DFM for the covariates (temperature and relative humidity) are also estimated, with $m = 24$ and $r = 2$ and 3, respectively.

Model 3C: This model has the same structure as Model 2C, but it uses actual values of the humidity and temperature covariates.

- As in Model 1C, $m = 24$, $r_0 = 2$, $k = 3$. $r_1 = 1$, $r_2 = 0$ and $r_3 = 1$. In this case, additional models for the covariates are not employed.

Remark 2. The fundamental difference between Models 1–3 and Models 1C - 3C is that the latter use information on the existence of clusters due to dependence between the time series. In this way, common factors that affect all series (r_0) and factors that are specific (r_i with $i \geq 1$) to each cluster are estimated. As observed, it may happen that some cluster does not have a specific factor associated with it.

Model 4: A linear regression model between generated energy and the covariates, using the predictions of the covariates that have been previously obtained, is used.

- The linear model considers the three variables: generated energy (response) and humidity and temperature (covariates) as univariate variables. Only, when the prediction for the covariates is obtained, the dimensions are $m = 24$ and $r = 2$ and 3, respectively.

Model 5: This model is identical in structure to Model 4 but real values of temperature and humidity covariates are used.

- As in the previous model, the generated energy (response) and humidity and temperature (covariates) are univariate variables. In this case, no additional models are used for the covariates.

Remark 3. Models 4 and 5 only take into account the relationship between the energy generated and the covariates at the same instant in time. This is a clear limitation considering the autocorrelation and cross-correlation structures illustrated in Figs. 3 and 5, respectively. The proposed models, 1 to 3 (1C to 3C), do take into account this temporal dependence, which allows the incorporation of the regency effect, that is, the dependence between the response variable and lags of the covariates.

Model 6: A Vector autoregressive model, VAR, as the one defined by (1) with the three considered variables, generated energy, temperature, and humidity.

- The dimensions of the vector time series (generated energy, temperature, and humidity) is 3×1. That is, $m = 3$.

Remark 4. Model 6 allows modeling the relationship between the three variables E_t, T_t and H_t (energy, temperature and relative humidity) with their delays. However, it has limitations if, as is the case, we want to carry out a parallel approach because the dimension of the model grows substantially. If we want to model the vector $\left(E_{t-1}, \ldots, E_{t-24}, T_{t-1}, \ldots, T_{t-24}, H_{t-1}, \ldots, H_{t-24}\right)$ using a VAR, we will have to estimate p matrices of dimension 72×72, where p is the order of the VAR model. Even if we only consider the vector $\left(E_{t-1}, E_{t-2}, \ldots, E_{t-24}\right)$, we would have to estimate 24×24 matrices. Dynamic factorial models such as models 1 to 3 (1C to 3C) are an effective response to the problem of the dimension of VAR models.

Models 1 to 3 (DFM) and 1C to 3C (DFMCS) are our methodological proposal. Schemes A.13–A.18 in the Appendix show a schematic representation of these six models. Models 4 to 6, that were proposed by the technology center, will be taken as benchmark.

In the models with group structure (Models 1C, 2C and 3C) three clusters ($k = 3$) have been found, see the dendrogram Fig. 9, in which the clustering of the time series corresponding to consecutive hours can be seen. It is important to remember that the vector time series, $\mathbf{y}_t$, is made up of the 24 time series of the measurements in each of the hours. When the time series of hour i is referred to, we are talking about the daily series formed by the observations measured at that specific hour. Fig. 9 shows that the time series corresponding to hours 1 to 12 (13 to 16, and 17 to 24) have a similar behavior and form three well-separated clusters. The cluster formed by the time series from hours 1 to 12 is denoted by C1, by C2 those from hours 13 to 17, and by C3 the series of the remaining hours. The specific factors are associated to the clusters with a larger number of observations. That is, the cluster conformed with the hours from 1 to 12 (C1) has a specific factor and the cluster conformed with the hours 17 to 24 (C3) has another specific factor.

In the next section, we also include the results using two automatic state-of-art time series models: (1) Trigonometric seasonality, Box–Cox transformation, ARMA errors, Trend and Seasonal components model (TBATS) proposed by Livera et al. (2011), and (2) Prophet proposed by Taylor and Letham (2017). These two methods have been used for prediction of electricity market prices and consumption by Karabiber

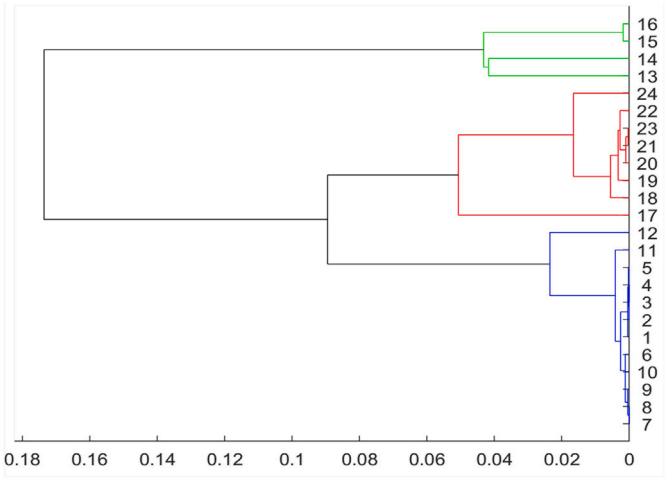

Fig. 9. Dendrogram of dynamic factor models with cluster structure.

and Xydis (2019) and Yildiz and Korkut (2024), respectively. In both models, the real values of the covariates (temperature and humidity) have been used, therefore the comparison must be made with the proposed models 3 and 3C.

3.4. Comparison of prediction results

In this section, the prediction results obtained from the nine models previously described above are presented, using the period between March 31, 2012, and April 14, 2013 as an out-of-sample testing period. The rolling windows procedure is employed with a window size of 90 days for model estimation, that is, all models are estimated using the 90 days prior to the two-day forecast horizon, with each step moving the window to the next day.

In Fig. 10, the daily MAPEs for the nine models are shown. The daily MAPEs are calculated with expression (7) using the n = 48 prediction errors for each day. Firstly, three mean absolute percentage errors with values higher than 20% stand out, which correspond to periods of unexpected drops in production. Fig. 11 shows the energy generated on each of those three days as well as the latest observations of its corresponding training set. In the three cases it is clearly observed that using historical values it is not feasible to predict these production drops, which may correspond to the shutdown of some production unit for one or two hours (days 71 and 201) or a longer period (day 22). Of course, if this information is known in advance, for example because it corresponds to a shutdown for unit maintenance, it can be incorporated into prediction models. It should be noted that all models fail in predicting these three days. It can be seen in the drop on day 22, which occurs at night and continues on day 23. In Fig. 10 we can see that when we incorporate the information from day 22, the prediction error for day 23 is much smaller because the model incorporates the detection of this drop. For the rest of the days, the percentage errors are smaller than 20%. In Fig. 12, the boxplots of the daily mean absolute percentage errors (MAPE) are presented for the remaining 376 days, i.e., omitting the three days with errors greater than 20% in order to do not distort the scale of the graph.

Furthermore, in Table 2 we show the deciles of the daily mean absolute percentage errors (MAPE) since it is important to focus not only on the central values but also on the upper deciles since these correspond to the days where the largest prediction errors occur.

The main conclusions drawn from Table 2 are as follows:

- As expected, models that use the actual values of the covariates obtain better results than analogous models that use predictions of the covariates.
- The best prediction results are obtained with Models 3 and 3C. It is worth noting that these models use the actual values of temperature and humidity, and therefore their results can be interpreted as the optimal value that could be obtained if these two covariates were predicted with high reliability.
- The comparison of Model 1, which does not use covariates, with Models 4 to 6, shows significant improvements from the 70th decile and in the mean error with the simpler model. This suggests that the factor models are capable of capturing the main dependence structures of the energy generation time series.
- The comparison between Model 1 and Model 2 indicates an improvement of between 5% and 15% in the deciles and 6% in the mean error. Obviously, this improvement is greater when compared to Model 3, increasing from 31% to 44% in the deciles and 25% in the mean error. This implies that the incorporation of meteorological covariates leads to a notable improvement in the predictions.
- The comparison of Model 2 with Model 4 reveals that the proposed model outperforms in almost all deciles (except the first decile) by between 5% and 49%, and by 22% in the mean error. It is noteworthy that there is an improvement in the upper deciles, with improvements of 29% and 49%, respectively. These upper deciles correspond to the largest prediction errors and therefore result in a greater deviation between the prediction and production and a greater penalty to the company.
- The conclusions of the previous point remain valid when comparing Model 3 with Model 5. In other words, the proposed model improves in all deciles except the first one and in the mean error of prediction.
- The comparison of Model 2 with Model 6 shows that the proposed model outperforms in all deciles by between 21% and 34%, and by 17% in the mean error. Obviously, this improvement is greater when compared to Model 3.
- It can be observed from the Table 2 that the cluster-structured models (Models 1C, 2C, and 3C) outperform the models that do not consider it (Models 1, 2, and 3, respectively). The percentage comparison between Model 1 and Model 1C shows an improvement ranging from 7% to 15% in the deciles and a 5% reduction in the mean error. For Model 2 and Model 2C, the improvement ranges from 6% to 14% in the deciles, and there is a 9% reduction in the mean error. Between Model 3 and Model 3C, the improvement increases to 37% to 46% in the deciles, and there is a 22% reduction in the mean error.
- The proposed models 3 and 3C obtains betters results than the TBATS and Prophet in all deciles and the mean. TBATS obtains competitive results but Prophet does not.

Finally, the boxplots in Fig. 12, indicate that the proposed models improve and have less dispersion in their results than the competitor models.

4. Conclusions

This paper introduces a novel modeling strategy utilizing Dynamic Factor Analysis (DFA) to predict cogeneration production, demonstrating superior performance compared to competitor models in terms of mean relative absolute error. The use of seasonal ARIMA models for modeling factors in DFA allows capturing the temporal dynamics of common factors and incorporating the cross-correlation structures between time series, resulting in a more flexible and robust approach compared to static principal component analysis. This flexibility allows DFA to handle non-stationary time series effectively, thereby improving its ability to analyze and predict complex time series. As a consequence, significant improvements have been achieved, with over 20% enhancement in the mean error and noteworthy advancements of close to 30% in the upper deciles, which correspond to the most costly errors for the company.

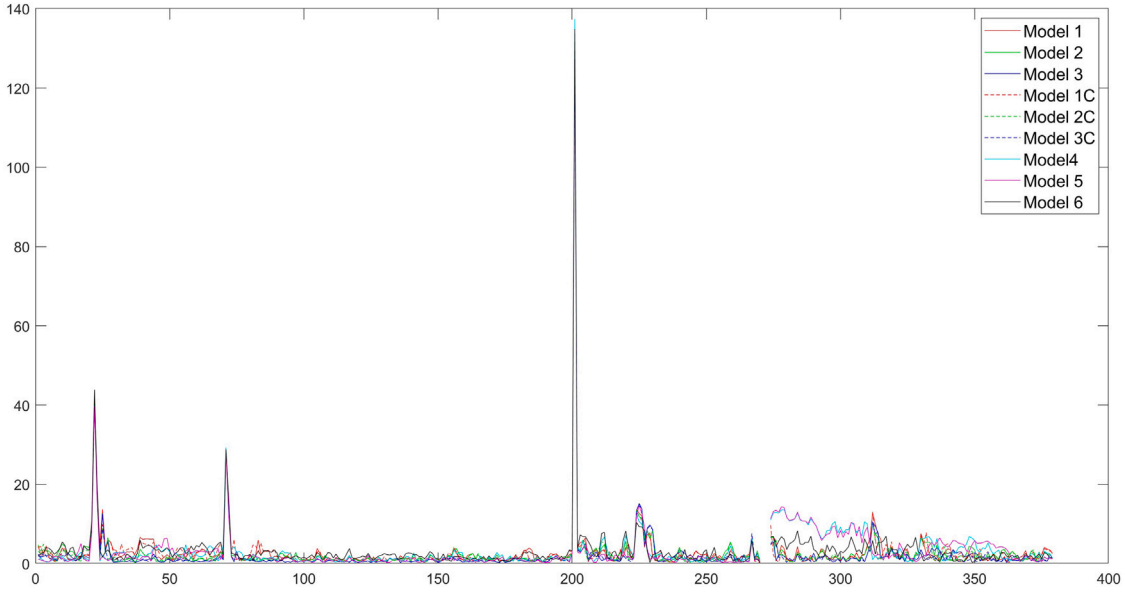

Fig. 10. Daily Mean Absolute Percentage Errors in the period 31/03/2012–12/04/2013. Models 1–6 (solid lines) and Models 1C–3C (dashed lines).

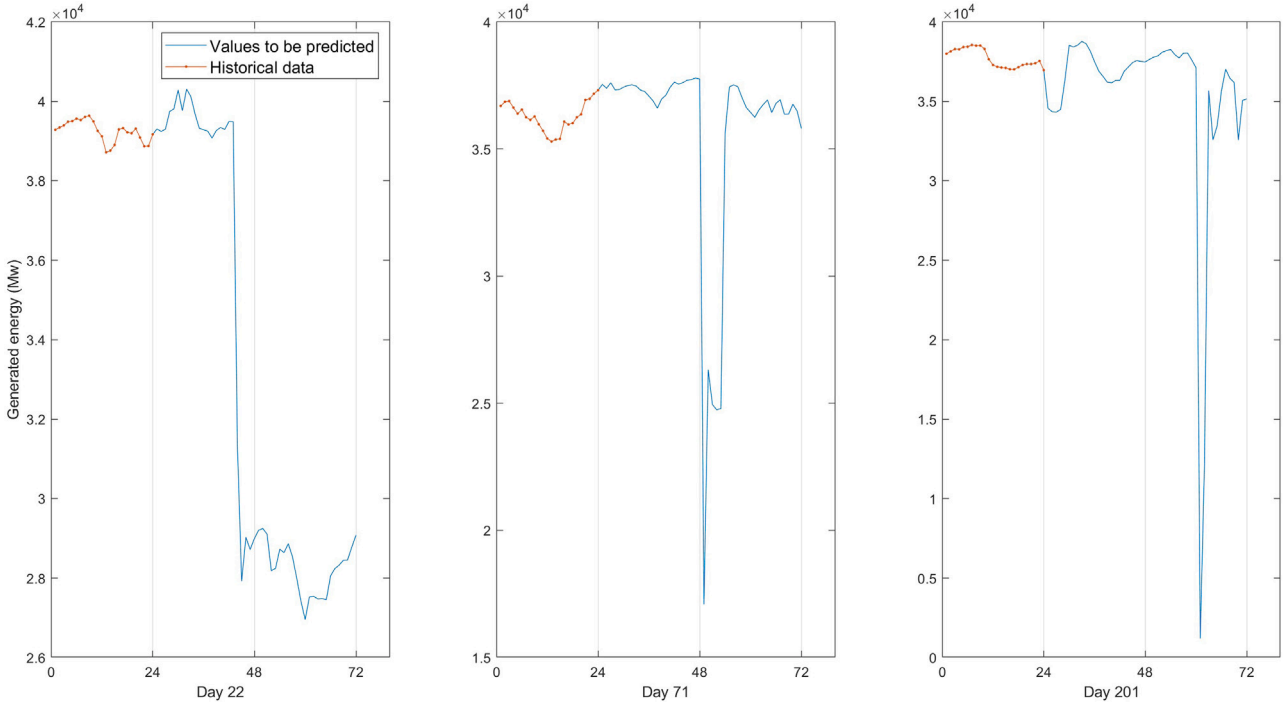

Fig. 11. Outlier days (22, 71, and 201) in relation to MAPE.

Table 2
Deciles and means of the daily mean absolute percentage errors.

	M1	M2	M3	M1C	M2C	M3C	M4	M5	M6	TBATS	Prophet
Decile 10	0.660	0.621	0.451	0.642	0.604	0.448	0.612	0.436	0.941	0.605	0.753
Decile 20	0.961	0.821	0.550	0.870	0.797	0.533	0.897	0.728	1.227	0.792	1.057
Decile 30	1.166	1.059	0.652	1.114	1.020	0.645	1.123	0.944	1.470	0.932	1.407
Decile 40	1.416	1.236	0.802	1.353	1.197	0.738	1.414	1.180	1.781	1.177	1.956
Median	1.711	1.447	0.977	1.624	1.387	0.872	1.741	1.418	2.217	1.423	3.038
Decile 60	1.970	1.847	1.250	2.043	1.720	1.095	2.216	1.811	2.588	1.687	4.197
Decile 70	2.374	2.247	1.628	2.760	2.044	1.277	2.786	2.714	3.111	2.028	5.507
Decile 80	3.215	2.846	1.943	3.238	2.707	1.666	4.037	4.642	3.854	2.547	8.342
Decile 90	4.644	4.206	2.946	4.630	3.844	2.676	8.278	7.666	5.392	3.770	16.792
Mean	2.849	2.663	2.136	2.827	2.543	1.994	3.438	3.283	3.244	2.787	5.852

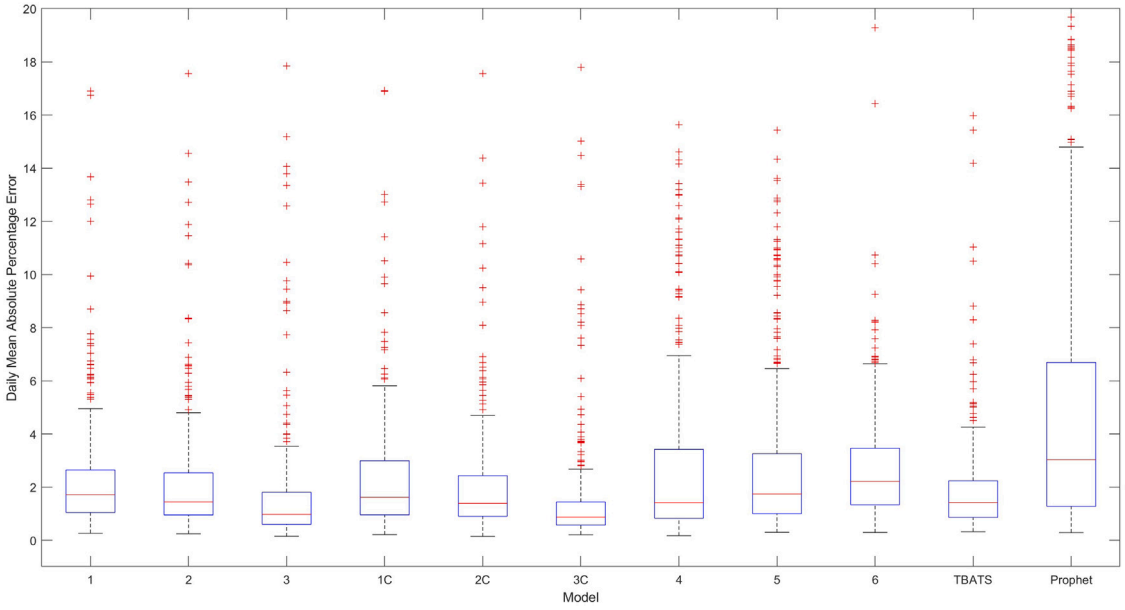

Fig. 12. Boxplots of daily MAPE obtained for the period 31/03/2012–12/04/2013.

Moreover, the incorporation of Dynamic Factor Models with Cluster Structure (DFMCS) in Models 1C, 2C, and 3C has further extended the model's capabilities and yielded significant improvements. DFMCS considers important features like heterogeneity, cluster structure, and the presence of multivariate outliers and level shifts, facilitating a more comprehensive and precise analysis. This innovative addition has demonstrated remarkable results, with percentage improvements ranging from 7% to 46% in the deciles and a substantial reduction of up to 25% in the mean error. Such advancements highlight the potential of DFMCS in enhancing cogeneration production predictions and its promising role in the energy forecasting domain.

The implementation of the local median interpolation method for handling missing values in the dataset has also contributed to the model's overall performance and reliability. It has enhanced data completeness and accuracy, further reinforcing the model's ability to provide robust predictions for cogeneration production.

Furthermore, the dynamic factor models used for predicting generated energy, with humidity and temperature as covariates, offer several advantages over machine learning and neural network models. These advantages include a clearer and simpler interpretation of the extracted factors, a greater ability to handle missing data and non-stationary time series, a lower risk of over-fitting, more effective dimensionality reduction, and multicollinearity elimination. Additionally, the ease of incorporating and adjusting various types of dynamic models to the data allows for enhanced adaptability and versatility in capturing complex relationships and dynamics within the energy production process.

Recent research has shown a wide range of approaches for forecasting energy consumption in various contexts, from individual buildings to entire regions. However, these methods face significant challenges in capturing the inherent complexity of energy consumption data, which can be influenced by various factors such as seasonal changes, long-term trends, and unexpected events. Although traditional techniques such as econometric models, machine learning, and neural networks can provide acceptable predictions under certain circumstances, they often struggle to adapt to rapidly changing energy consumption patterns or to capture nonlinear relationships in the data. Unlike the reviewed models, which often require specific assumptions or are limited by their ability to handle complex relationships, DFA and DFMCS can efficiently integrate multiple sources of information and dynamically adapt their latent factors to improve prediction accuracy and robustness. Moreover, DFMs allow for long-term forecasts, in this specific work up to two days (48 h) ahead, with greater accuracy than the other models analyzed.

In summary, the combination of DFA and DFMCS has proven to be a valuable modeling approach for predicting cogeneration production, enabling energy companies to optimize their production, improve their financial performance, and reduce the environmental impact of energy production.

An aspect that has not been addressed in this paper and that will be the subject of future research is the incorporation of uncertainty in predictions. Recent reviews can be found in Hong and Fan (2016) and Lin et al. (2023). An alternative that has been considered in ARIMA models has been the use of the bootstrap, not only to incorporate the uncertainty due to the estimation of the parameters (see Pascual et al., 2004) but also to the selection of the models (see Alonso et al., 2006).

Abbreviations

The following abbreviations are used in this manuscript:

ARIMA	Autoregressive Integrated Moving Average model
ARIMAX	Autoregressive Integrated Moving Average model with eXogenous inputs
CHP	Cogeneration
DFA	Dynamic Factor Analysis
DFM	Dynamic Factor Model
DFMCS	Dynamic Factor Model with Cluster Structure
MAPE	Mean Absolute Percentage Error
PCA	Principal Component Analysis
SEATS	Signal Extraction in ARIMA Time Series
TBATS	Trigonometric seasonality, Box–Cox transformation, ARMA errors, Trend and Seasonal components model
TRAMO	Time series Regression with ARIMA noise, Missing values, and Outliers
VAR	Vector AutoRegression model

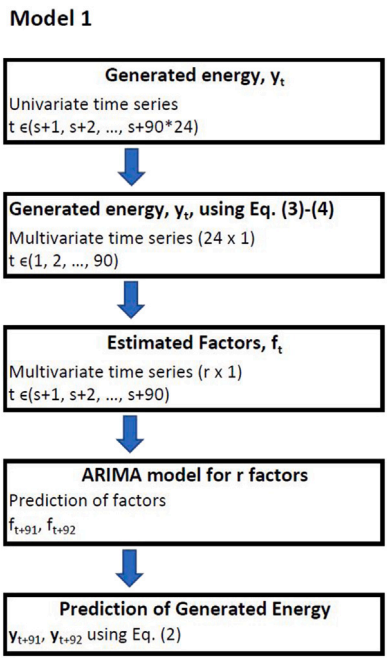

Fig. A.13. Schematic representation of Model 1.

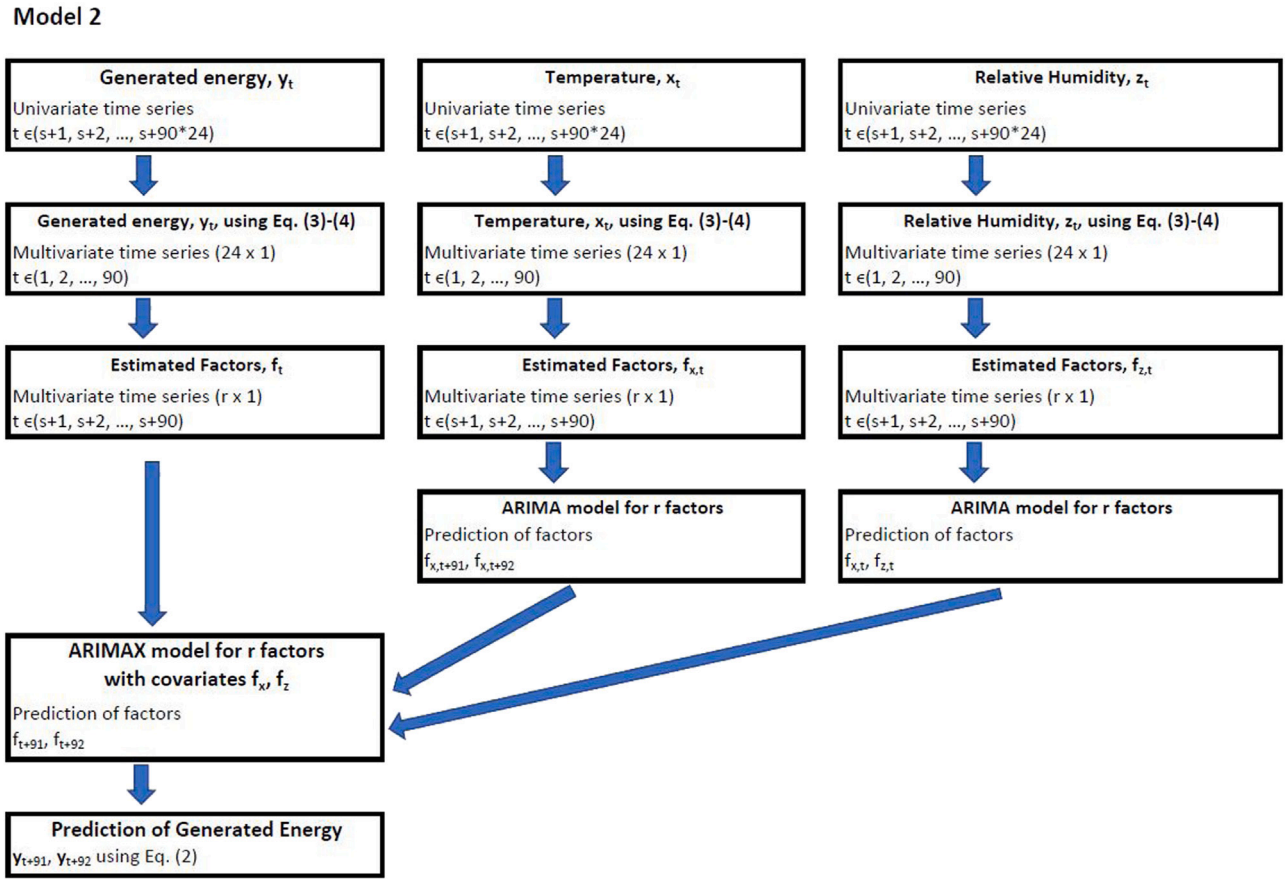

Fig. A.14. Schematic representation of Model 2.

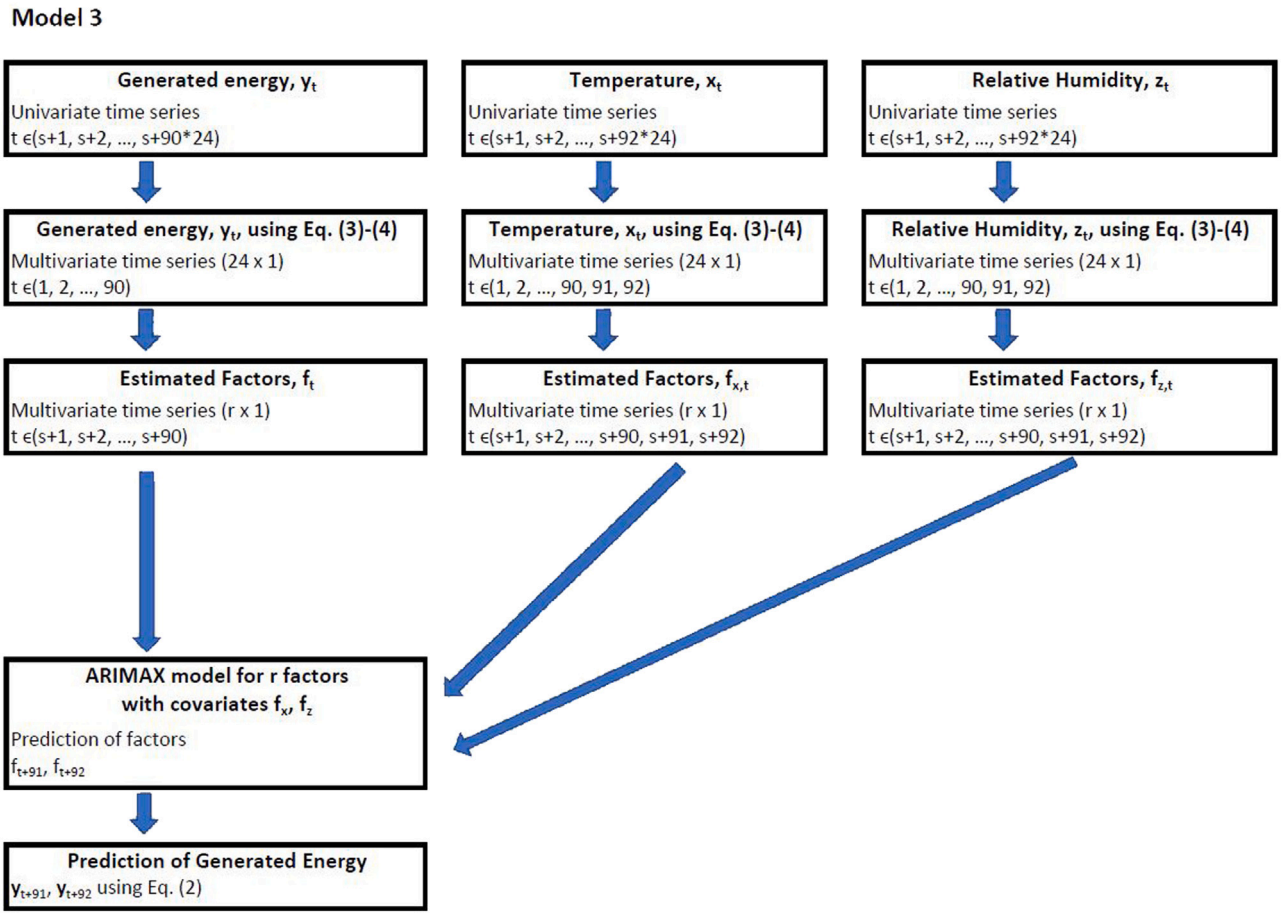

Fig. A.15. Schematic representation of Model 3.

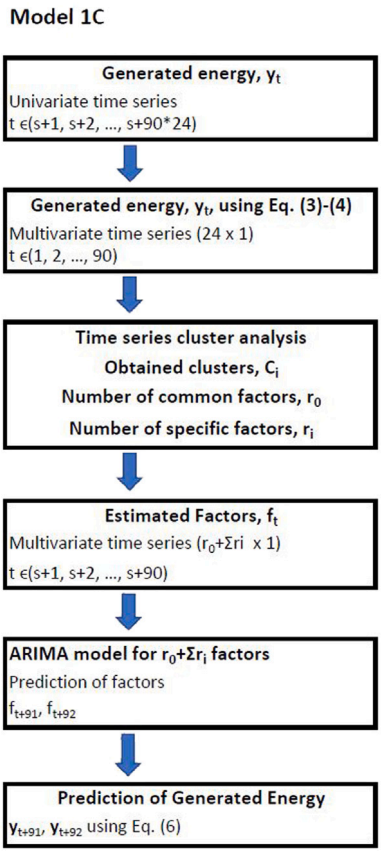

Fig. A.16. Schematic representation of Model 1C.

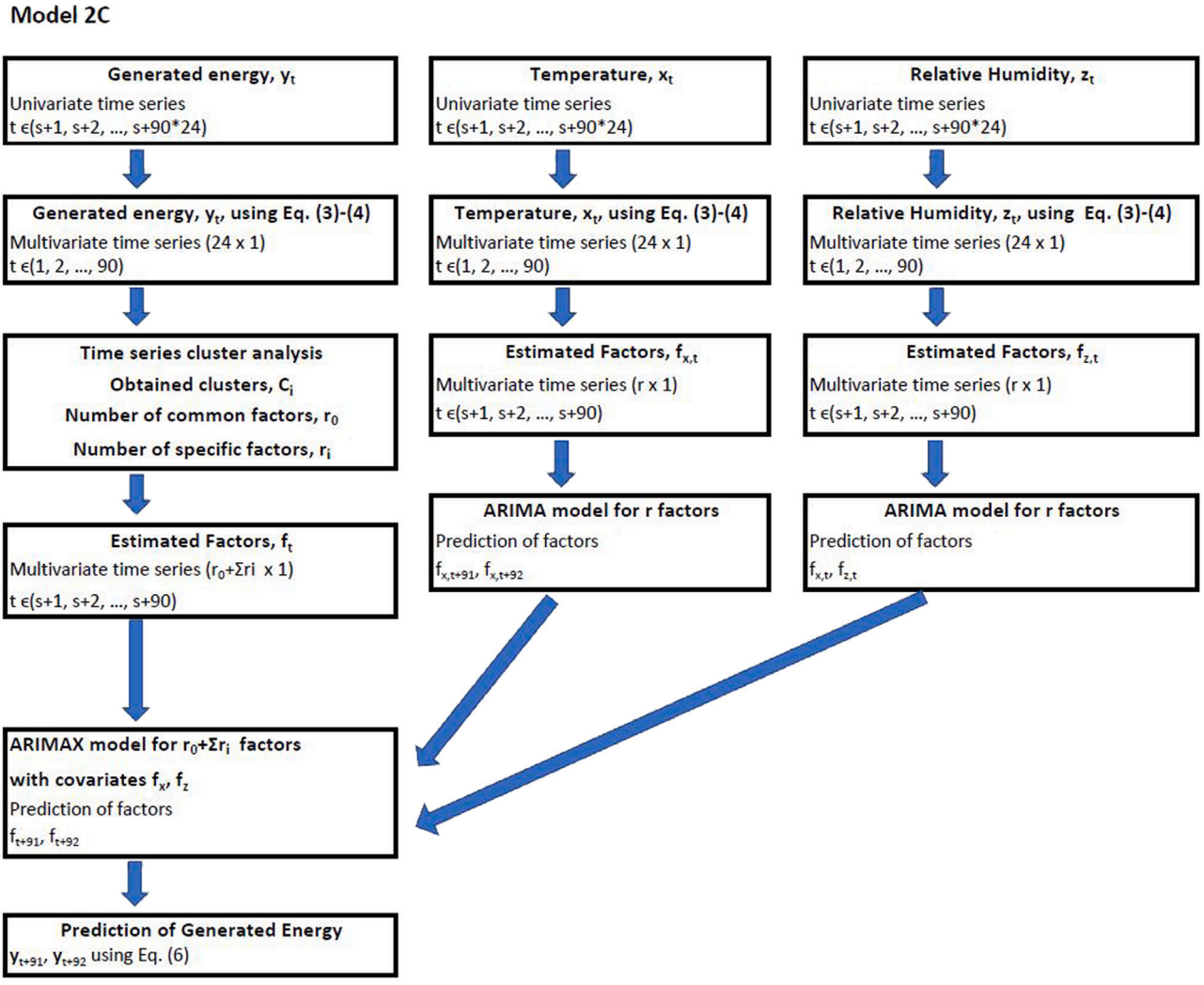

Fig. A.17. Schematic representation of Model 2C.

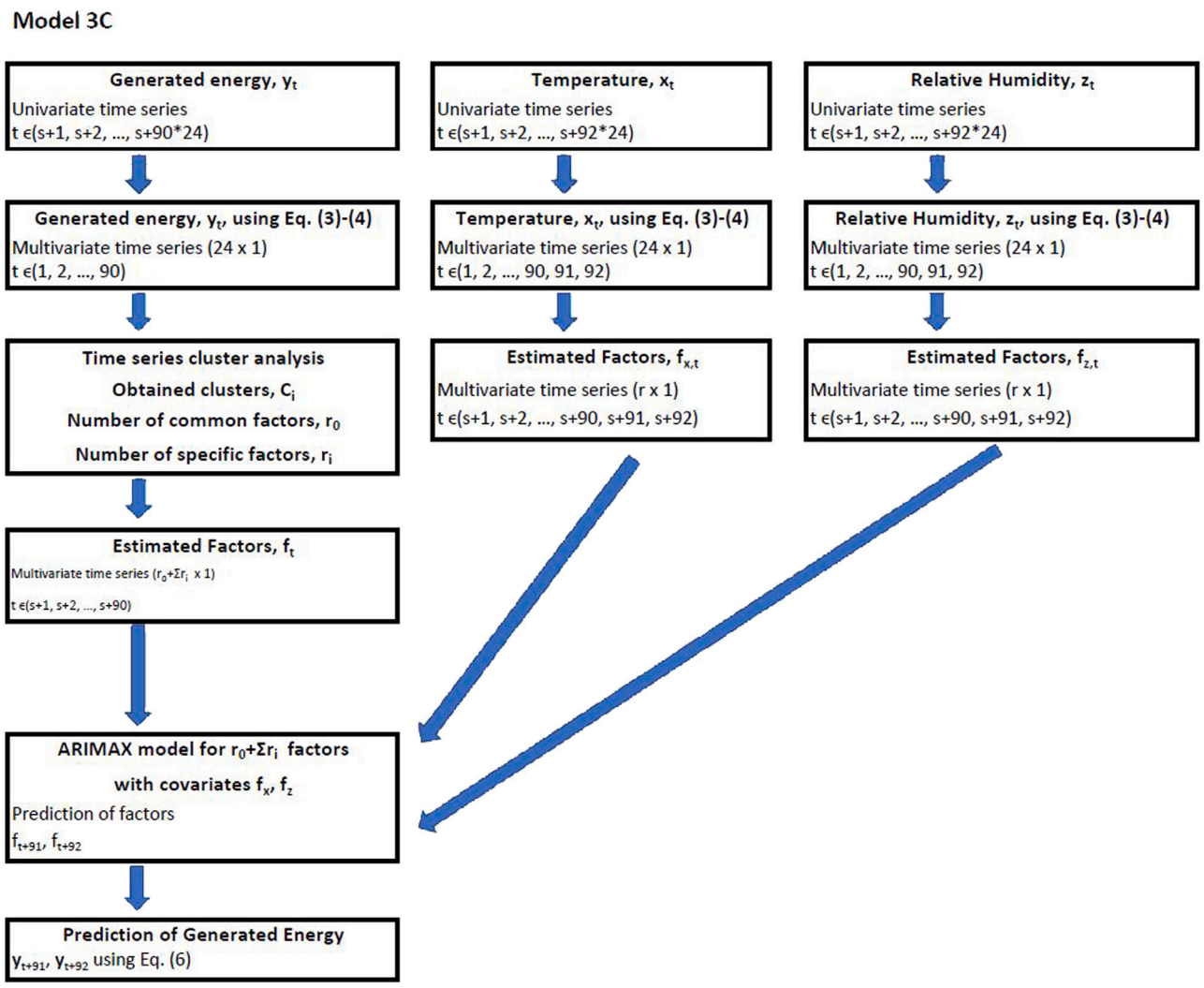

Fig. A.18. Schematic representation of Model 3C.

CRediT authorship contribution statement

Andrés M. Alonso: Writing – review & editing, Software, Methodology, Conceptualization. **A.E. Sipols:** Software, Methodology, Conceptualization. **M. Teresa Santos-Martín:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research was funded by Ministry of Science and Innovaton projects:

PID2019-108311GB-I00

PID2022-138114NB-I00

PID2021-125211OB-I00

and by the Junta de Castlla y León project:

SA212P23.

Appendix

The appendix shows a schematic representation of the six forecasting models described in Section 3.3 (see Figs. A.13–A.18).

References

Alonso, A. M., Bastos, G., & García-Martos, C. (2016). Electricity price forecasting by averaging dynamic factor models. *Energies*, *9*(8), 600. http://dx.doi.org/10.3390/en9080600.

Alonso, A. M., Galeano, P., & Peña, D. (2020). A robust procedure to build dynamic factor models with cluster structure. *Journal of Economics*, *216*(1), 35–52. http://dx.doi.org/10.1016/j.jeconom.2020.01.004.

Alonso, A. M., Peña, D., & Romo, J. (2006). Introducing model uncertainty by moving blocks bootstrap. *Statistical Papers*, *47*(2), 167–179. http://dx.doi.org/10.1111/j.1467-9892.2004.01713.x.

Ando, T., & Bai, J. (2017). Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association*, *112*(519), 1182–1198. http://dx.doi.org/10.1080/01621459.2016.1195743.

Bedi, J., & Toshniwal, D. (2019). Deep learning framework to forecast electricity demand. *Applied Energy*, *238*, 1312–1326. http://dx.doi.org/10.1016/j.apenergy.2019.01.113.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Bujalski, M., & Madejski, P. (2021). Forecasting of heat production in combined heat and power plants using generalized additive models. *Energies*, *14*(8), 2331. http://dx.doi.org/10.3390/en14082331.

Chatfield, C. (2013). *The analysis of time series: theory and practice*. Springer.

Chaturvedi, S., Rajasekar, E., Natarajan, S., & McCullen, N. (2022). A comparative assessment of SARIMA, LSTM RNN and Fb Prophet models to forecast total and peak monthly energy demand for India. *Energy Policy*, *168*, Article 113097. http://dx.doi.org/10.1016/j.enpol.2022.113097.

Chung, W. H., Gu, Y. H., & Yoo, S. J. (2022). District heater load forecasting based on machine learning and parallel CNN-LSTM attention. *Energy*, *246*, Article 123350. http://dx.doi.org/10.1016/j.energy.2022.123350.

Deb, C., Zhang, F., Yang, J., Lee, S. E., & Shah, K. W. (2017). A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, *74*, 902–924. http://dx.doi.org/10.1016/j.rser.2017.02.085.

Deng, S., Zhang, J., Zhang, C., Luo, M., Ni, M., Li, Y., & Zeng, T. (2022). Prediction and optimization of gas distribution quality for high-temperature PEMFC based on data-driven surrogate model. *Applied Energy*, *327*, Article 120000. http://dx.doi.org/10.1016/j.apenergy.2022.120000.

Dordonnat, V., Koopman, S. J., & Ooms, M. (2012). Dynamic factors in periodic time-varying regressions with an application to hourly electricity load modelling. *Computational Statistics & Data Analysis*, *56*(11), 3134–3152. http://dx.doi.org/10.1016/j.csda.2011.04.002.

Ebrahimi-Moghadam, A., Moghadam, A. J., Farzaneh-Gord, M., & Arabkoohsar, A. (2021). Performance investigation of a novel hybrid system for simultaneous production of cooling, heating, and electricity. *Sustainable Energy Technologies and Assessments*, *43*, Article 100931. http://dx.doi.org/10.1016/j.seta.2020.100931.

Enders, W. (2004). *Applied econometric time series*. John Willey and Son.

Fan, C., Sun, Y., Zhao, Y., Song, M., & Wang, J. (2019). Deep learning-based feature engineering methods for improved building energy prediction. *Applied Energy*, *240*, 35–45. http://dx.doi.org/10.1016/j.apenergy.2019.02.052.

García-Martos, C., Rodríguez, J., & Sánchez, M. (2012). Forecasting electricity prices by extracting dynamic common factors: application to the Iberian market. *IET Generation, Transmission & Distribution*, *6*(1), 11–20. http://dx.doi.org/10.1049/iet-gtd.2011.0009.

Gómez, V., & Maravall Herrero, A. (1998). *Guide for using the programs TRAMO and SEATS: beta version: December 1997*. Banco de España. Servicio de Estudios.

He, H., Chen, L., & Wang, S. (2023). Flight short-term booking demand forecasting based on a long short-term memory network. *Computers & Industrial Engineering*, *186*, Article 109707. http://dx.doi.org/10.1016/j.cie.2023.109707.

He, Y., Guo, S., Zhou, J., Wu, F., Huang, J., & Pei, H. (2021). The many-objective optimal design of renewable energy cogeneration system. *Energy*, *234*, Article 121244. http://dx.doi.org/10.1016/j.energy.2021.121244.

Hong, T., & Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, *32*(3), 914–938. http://dx.doi.org/10.1016/j.ijforecast.2015.11.011.

Ifaei, P., Nazari-Heris, M., Charmchi, A. S. T., Asadi, S., & Yoo, C. (2023). Sustainable energies and machine learning: An organized review of recent applications and challenges. *Energy*, *266*, Article 126432. http://dx.doi.org/10.1016/j.energy.2022.126432.

Ikeda, S., & Nagai, T. (2021). A novel optimization method combining metaheuristics and machine learning for daily optimal operations in building energy and storage systems. *Applied Energy*, *289*, Article 116716. http://dx.doi.org/10.1016/j.apenergy.2021.116716.

Karabiber, O. A., & Xydis, G. (2019). Electricity price forecasting in the Danish day-ahead market using the TBATS, ANN and ARIMA methods. *Energies*, *12*, 928. http://dx.doi.org/10.3390/en12050928.

Klyuev, R. V., Morgoev, I. D., Morgoeva, A. D., Gavrina, O. A., Martyushev, N. V., Efremenkov, E. A., & Mengxu, Q. (2022). Methods of forecasting electric energy consumption: A literature review. *Energies*, *15*(23), 8919. http://dx.doi.org/10.3390/en15238919.

Lee, Y., & Kang, S. (2024). Dynamic ensemble of regression neural networks based on predictive uncertainty. *Computers & Industrial Engineering*, Article 110011. http://dx.doi.org/10.1016/j.cie.2024.110011.

Lin, F., Zhang, Y., & Wang, J. (2023). Recent advances in intra-hour solar forecasting: A review of ground-based sky image methods. *International Journal of Forecasting*, *39*(1), 244–265. http://dx.doi.org/10.1016/j.ijforecast.2021.11.002.

Livera, A. M. D., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, *106*(496), 1513–1527. http://dx.doi.org/10.1198/jasa.2011.tm09771.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer.

Lütkepohl, H. (2009). *Econometric analysis with vector autoregressive models* (pp. 281–319). Wiley Online Library, http://dx.doi.org/10.1002/9780470748916.ch8.

Maravall, A., López-Pavón, R., & Pérez-Cañete, D. (2015). Reliability of the automatic identification of ARIMA models in program TRAMO. In J. Beran, Y. Feng, & H. Hebbel (Eds.), *Empirical economic and financial research: theory, methods and practice* (pp. 105–122). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-03122-4_7.

Mezzi, R., Yousfi-Steiner, N., Péra, M. C., Hissel, D., & Larger, L. (2021). An Echo State Network for fuel cell lifetime prediction under a dynamic micro-cogeneration load profile. *Applied Energy*, *283*, Article 116297. http://dx.doi.org/10.1016/j.apenergy.2020.116297.

Miroshnyk, V., Shymaniuk, P., & Sychova, V. (2021). Short term renewable energy forecasting with deep learning neural networks. In *Power syst. res. and operation: selected problems* (pp. 121–142). Springer, http://dx.doi.org/10.1007/978-3-030-82926-1_6.

Nepal, B., Yamaha, M., Yokoe, A., & Yamaji, T. (2020). Electricity load forecasting using clustering and ARIMA model for energy management in buildings. *Japan Architectural Review*, *3*(1), 62–76. http://dx.doi.org/10.1002/2475-8876.12135.

Ni, Z., Zhang, C., Karlsson, M., & Gong, S. (2024). A study of deep learning-based multi-horizon building energy forecasting. *Energy and Buildings*, *303*, Article 113810. http://dx.doi.org/10.1016/j.enbuild.2023.113810.

Pascual, L., Romo, J., & Ruiz, E. (2004). Bootstrap predictive inference for ARIMA processes. *Journal of Time Series Analysis*, *25*(4), 449–465. http://dx.doi.org/10.1111/j.1467-9892.2004.01713.x.

Peña, D., & Box, G. E. (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, *82*(399), 836–843. http://dx.doi.org/10.1080/01621459.1987.10478506.

Pu, Z., & Kalnay, E. (2019). Numerical weather prediction basics: Models, numerical methods, and data assimilation. In Q. Duan, F. Pappenberger, A. Wood, H. L. Cloke, & J. C. Schaake (Eds.), *Handbook of hydrometeorological ensemble forecasting* (pp. 67–97). Berlin, Heidelberg: Springer Berlin Heidelberg, http://dx.doi.org/10.1007/978-3-642-39925-1_11.

Ramanathan, R., Engle, R., Granger, C. W., Vahid-Araghi, F., & Brace, C. (1997). Short-run forecasts of electricity loads and peaks. *International Journal of Forecasting*, *13*(2), 161–174. http://dx.doi.org/10.1016/S0169-2070(97)00015-0.

Runge, J., & Saloux, E. (2023). A comparison of prediction and forecasting artificial intelligence models to estimate the future energy demand in a district heating system. *Energy*, *269*, Article 126661. http://dx.doi.org/10.1016/j.energy.2023.126661.

Sarathkumar, T. V., & Goswami, A. K. (2022). Renewable energy resources forecasting model for virtual power plant in the deregulated electricity market using machine learning. In *Inter. conf. power electr., smart grid, and renew. energy* (pp. 1–6). PESGRE, http://dx.doi.org/10.1109/PESGRE52268.2022.9715958.

Shaikh, A. K., Nazir, A., Khalique, N., Shah, A. S., & Adhikari, N. (2023). A new approach to seasonal energy consumption forecasting using temporal convolutional networks. *Results in Engineering*, *19*, Article 101296. http://dx.doi.org/10.1016/j.rineng.2023.101296.

Stock, J. H., & Watson, M. W. (1988). Testing for common trends. *Journal of the American Statistical Association*, *83*(404), 1097–1107. http://dx.doi.org/10.1080/01621459.1988.10478707.

Taşcı, B., Omar, A., & Ayvaz, S. (2023). Remaining useful lifetime prediction for predictive maintenance in manufacturing. *Computers & Industrial Engineering*, *184*, Article 109566. http://dx.doi.org/10.1016/j.cie.2023.109566.

Taylor, S., & Letham, B. (2017). Forecasting at scale. *PeerJ Preprints*, *5*, e3190v2. http://dx.doi.org/10.7287/peerj.preprints.3190v2.

Teichgraeber, H., & Brandt, A. R. (2019). Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison. *Applied Energy*, *239*, 1283–1293. http://dx.doi.org/10.1016/j.apenergy.2019.02.012.

Vialetto, G., & Noro, M. (2020). An innovative approach to design cogeneration systems based on big data analysis and use of clustering methods. *Energy Conversion and Management*, *214*, Article 112901. http://dx.doi.org/10.1016/j.enconman.2020.112901.

Wang, P., Liu, B., & Hong, T. (2016). Electric load forecasting with recency effect: A big data approach. *International Journal of Forecasting*, *32*(3), 585–597. http://dx.doi.org/10.1016/j.ijforecast.2015.09.006.

Weber, T., Sossenheimer, J., Schäfer, S., Ott, M., Walther, J., & Abele, E. (2019). Machine learning based system identification tool for data-based energy and resource modeling and simulation. *Procedia CIRP*, *80*, 683–688. http://dx.doi.org/10.1016/j.procir.2018.12.021.

Yan, Z., Zhou, X., Zhang, Q., Du, R., & Ren, J. (2024). Predicting multi-subsequent events and actors in public health emergencies: An event-based knowledge graph approach. *Computers & Industrial Engineering*, *187*, Article 109852. http://dx.doi.org/10.1016/j.cie.2023.109852.

Yildiz, U., & Korkut, S. O. (2024). Electricity consumption forecasting using the prophet model in industry: A case study. In F. P. García Márquez, A. Jamil, I. S. Ramirez, S. Eken, & A. A. Hameed (Eds.), *Computing, internet of things and data analytics* (pp. 102–111). Cham: Springer Nature Switzerland.