Leveraging spatio-temporal autocorrelation to improve the forecasting of the energy consumption in smart grids

Annunziata D'Aversa[1][0000−0003−1791−5998], Stefano Polimena[1][0000−0003−2181−5631], Gianvito Pio[1,2][0000−0003−2520−3616], and Michelangelo Ceci[1,2,3][0000−0002−6690−7583]

Dept. of Computer Science, University of Bari Aldo Moro, Bari, Italy[1]
Big Data Lab, CINI Consortium, Rome, Italy[2]
Dept. of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia[3]
{annunziata.daversa,stefano.polimena,gianvito.pio,
michelangelo.ceci}@uniba.it

Abstract. Smart grids are networks that distribute electricity by relying on advanced communication technologies, sensor measurements, and predictive methods, to quickly adapt the network behavior to different possible scenarios. In this context, the adoption of machine learning approaches to forecast the customer energy consumption is essential to optimize network planning operations, avoid unnecessary energy production, and minimize power shortages. However, classical forecasting methods are not able to take into account spatial and temporal autocorrelation phenomena, naturally introduced by the spatial proximity of consumers, and by the seasonality of the energy consumption trends.
In this paper, we investigate the adoption of several solutions to take into account spatio-temporal autocorrelation phenomena. Specifically, we investigate the contribution provided by the explicit representation of temporal information related to historical measurements using multiple strategies, as well as that of simultaneously predicting multiple future consumption measurements in a multi-step predictive setting. Finally, we investigate the effectiveness of injecting descriptive features to make the learning methods aware of the spatial closeness among the consumers. The experimental evaluation performed on a real-world electrical network demonstrated the positive contribution of making the models aware of spatio-temporal autocorrelation phenomena, and proved the overall superiority of models based on the multi-step predictive setting.

Keywords: Energy forecasting · Multi-step prediction · Spatio-temporal autocorrelation

1 Introduction

The infrastructures for the energy distribution are continuously subject to evolutions, mainly because of the generally increasing energy demand, as well as

of the introduction of new technologies, such as renewable power plants and car charging stations. The need of managing complex scenarios led to the definition of the so-called *smart grids*, that are distribution networks that exploit sensor measurements, advanced communication technologies and predictive components, to quickly adapt the network behavior to multiple possible situations. In this context, the accurate forecasting of the customer energy consumption is fundamental, not only to optimize the planning of network maintenance operations over the long term, but also to properly tune the production of energy from fossil fuel power stations. Indeed, producing energy from fossil sources generally leads to high CO_2 emissions, and the overproduction may also lead to the need of additional resources for storage. On the other hand, the underestimation of the energy consumption may compromise the system reliability, since an excessive demand could easily degenerate into a blackout. For these reasons, it is of paramount importance to predict the energy consumption in the network.

Machine learning methods can fruitfully be adopted to support this task, since they are able to exploit historical data, temporal trends and other consumer characteristics to build accurate predictive models. In general, the temporal dimension plays a central role for this task. Indeed, we can expect to observe cyclical behaviors, for example, along the months of each year (i.e., a generally higher consumption during summer and winter, mainly due to heating/cooling systems, rather than during spring and autumn).

The temporal dimension can generally introduce autocorrelation phenomena, known as the correlation of a signal with a *delayed copy* of itself as a function of delay, or the similarity between observations as a function of the time lag between them [4]. Analogously, the spatial closeness can influence the measurements: the Tobler's first law of geography [17] states that "everything is related to everything else, but near things are more related than distant things". In this specific context, spatially close consumers may exhibit a similar behavior, mainly because they live in similar climatic conditions. Although considering temporal and spatial autocorrelation phenomena should generally lead to a higher accuracy of the learned models [15], they have not yet been fully exploited in the context of the prediction of the energy consumption. Indeed, in the literature we can find only few works that investigated their contribution for the forecasting of the energy consumption, which are based on classical ARIMA models [6, 12]. On the other hand, their positive effect on the accuracy of the learned predictive models has been observed in the context of the energy production from photovoltaic power plants [5]. However, the challenges arising while aiming to predict the energy production and the energy consumption are different: while the former task is much more dependent on physical factors, such as weather conditions, in the latter, the prediction is mainly dependent on the behavior of consumers. Therefore, it is expected that the temporal dimension is more influential on the prediction of the energy consumption than for the prediction of the energy production.

In this paper, we propose a method for the forecasting of the monthly energy consumption of the consumers of a smart grid on a yearly horizon. The proposed

approach is able to properly capture and model both temporal and spatial autocorrelation phenomena. Different strategies are proposed for both the temporal and the spatial dimensions, each of which is able to properly model specific temporal/spatial characteristics and relationships among different measurements. Finally, we investigate the possibility to predict the 12 monthly measurements of the considered yearly horizon simultaneously, in a multi-step predictive setting, that, as we will emphasize in Sec. 2, is able to implicitly model the temporal relationships among the measurements at different time points, for both descriptive and target variables.

The rest of the paper is organized as follows. In Sec. 2, we briefly discuss existing related work. In Sec. 3, we describe the proposed approach for the forecasting of the energy consumption in smart grids, taking into account both temporal and spatial autocorrelation phenomena. In Sec. 4 we describe our experiments on a real-world energy distribution network. Finally, in Sec. 5, we draw some conclusions and outline possible future work.

2 Related Work

In the literature, we can find several works that propose methods for the prediction of the energy consumption, at different spatial and temporal scales: from high and very localized geographical resolutions (e.g., hourly measurements of a single sensor) to coarser temporal resolutions (e.g., days, months, years) and/or covering a large geographic area (e.g., a region or a country). Existing approaches can also be categorized as *single-step* methods, that aim to predict the value of a target attribute for a single future time step, and *multi-step* methods, that aim to predict the value of a target attribute for multiple steps ahead. In [16], the authors described different strategies that can be adopted to solve the latter task, including *recursive*, *direct* and *Multi-Input Multi-Output (MIMO)* strategies. The *recursive* strategy exploits an approach based on self learning, that iterates a single-step ahead predictive model to obtain the desired forecasts: after estimating the next value of the sequence, it is fed back as a descriptive variable for the subsequent prediction. The *direct* strategy is based on learning a set of independent predictive models, where the i-th model is able to return a prediction for the i-th time points in the future. Note that both *recursive* and *direct* strategies are actually single-step approaches that are applied multiple times to obtain a multi-step ahead prediction. On the other hand, the *MIMO* strategy aims to learn one global model that returns a vector of predictions, also possibly taking into account the existence of dependencies between future values, that in principle may be beneficial in terms of forecasting accuracy [3].

In [2], the authors proposed a deep learning architecture to forecast the customer energy consumption for the next month, using the measurements of the previous 12 months and other information such as the target month and the category of the customer (e.g., residential, business, etc.). Among the considered deep learning models, LSTM achieved the lowest mean absolute error.

In [18] the authors compared the performance of different methods, such as Linear Regression, Regression Trees and Multivariate Adaptive Regression Spline (MARS), for the prediction of the next month energy consumption using climate data and the characteristics of the buildings (e.g., size of living area, number of rooms, etc.). The authors also aggregated the individual consumptions to predict the monthly consumption for groups of buildings. Results showed that MARS was the best model for individual households, while regression trees outperformed the competitors for the prediction of the consumption of the groups.

In [10], the authors adopted the *direct* strategy to predict the electric load 10 days ahead using ARIMA and LSTM. The models were evaluated on three electrical networks and the results showed a general superiority of LSTM.

Despite several studies have been proposed for energy consumption forecasting, only a few of them investigated the possible contribution coming from spatial and temporal autocorrelation phenomena. An attempt in this direction has been done in [6, 12], where the authors considered spatial autocorrelation phenomena for the forecasting of the regional electricity consumption. In these works, a spatial ARMA model (SAR-ARMA) and a spatial ARIMA model (ARIMA-Sp) were proposed. However, auto-regressive approaches usually train a model based on the target variable only, and are not able to take into account additional features and possible dependencies between them and the target variable.

In [9], the authors proposed a deep neural network, called LSTNet, which combines convolutional neural networks to capture short-term patterns and LSTM or GRU for long-term patterns. To overcome the issue caused by the vanishing gradient, which affects the possibility to properly capture long-term interdependencies, the authors proposed the introduction of a recurrent-skip layer or an attention mechanism. Similarly, in [14], the authors proposed TPA-LSTM, an attention-recurrent neural network that allows the model to learn interdependencies among multiple variables across all previous time-steps.

The consideration of the spatial and of the temporal dimensions gained a general interest for other tasks related to time-series forecasting, even if not specifically focused on the prediction of the energy consumption. In particular, neural network architectures that simultaneously consider both temporal and spatial dimensions have been recently proposed. A relevant example is Graph WaveNet [19], a spatio-temporal graph convolutional network for multi-step forecasting, tailored for the prediction of traffic conditions at different locations. It uses dilated convolution networks to capture temporal dependencies and a self-adaptive adjacency matrix to capture spatial correlations. Another relevant example applied in the same domain is GMAN [20], which exploits a graph multi-attention network, with spatial and temporal attention mechanisms. Since it can be considered as one of the most recent approaches for multi-step prediction, that also consider spatio-temporal aspects, it will be considered as a state-of-the-art competitor in our experimental evaluation (see Sec. 4).

3 The proposed method

In this section, we describe our approach to forecast the monthly energy consumption of consumers on a yearly horizon. Therefore, the goal is to predict, for each consumer, 12 energy consumption values, i.e., one for each month of the subsequent year. As mentioned in Sec. 1, predicting such values is useful for planning network maintenance operations, as well as for tuning the energy production from fossil sources.

In the following subsections, we report the details of the proposed strategies to take into account the temporal and the spatial autocorrelation phenomena. After properly representing the temporal and the spatial dimensions, different standard regression models can be learned on top. At the end of the following subsection, we also briefly introduce the considered regressors and their extension to the multi-step predictive setting proposed in this paper.

3.1 Modeling the temporal autocorrelation

We propose different strategies to take into account the temporal autocorrelation, exploiting historical data about consumptions. We investigate two forecasting settings, namely, single-step (SS), where the 12 predictions are obtained by a *recursive* approach, and multi-step (MS), falling in the *MIMO* category, which goal is that of learning a global predictive model that returns the whole vector of 12 predictions. More formally, considering a time series of length w of energy consumptions for the consumer c, the SS setting consists in the exploitation of the historical measurements up to the time-step t-1 to predict the next time-step $y_{c,t}$. Through the recursive strategy, the predicted value $y_{c,t}$ is considered as a real measurement for the forecast of the energy consumption $y_{c,t+1}$, and so on up to predict $y_{c,t+11}$ (see the left part of Fig. 1).

Note that the adopted *recursive* strategy exhibits both advantages and disadvantages with respect to the *direct* strategy. Among the strong points, we can mention that the number of training instances increases (roughly by a factor of w), thanks to the fact that the measurement at a given month is considered multiple times, in different positions of the w-dimensional training time series (see, for example, the measurement related to Dec 2018 in the left part of Fig. 1). On the other hand, this aspect introduces the disadvantage of losing the temporal semantics of each descriptive feature, namely, each feature does not represent the same month of the year for all the training instances. This means that the model learned in this setting cannot easily detect and exploit seasonality phenomena. Another disadvantage is that, since it relies on a self-training approach, forecasting errors at the initial time-steps may be propagated to subsequent time-steps [13]. In order to alleviate the first issue, keeping the advantages of the recursive strategy, we explicitly represent the temporal information through additional features. In this respect, we propose two alternative settings:

– **SS-DT** (Described Target time-step), that introduces two additional descriptive features, namely the year j_t and the month m_t of the target value to predict $y_{c,t}$;

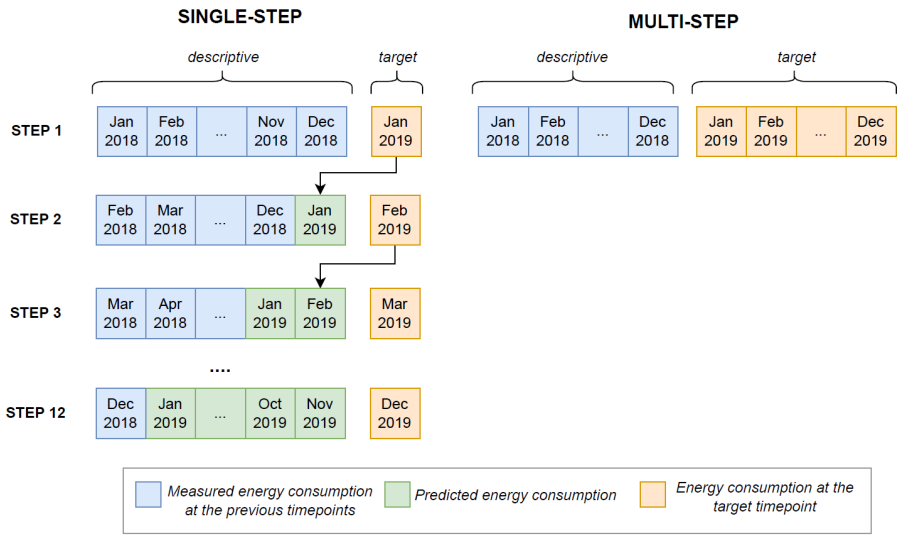

Fig. 1. A graphical representation of the single-step (SS) and multi-step (MS) learning settings. In the SS setting, the prediction for the i-th step is added to the descriptive variables for the prediction of the $(i+1)$-th step, while in the MS setting a global method able to simultaneously predict the value for all the 12 steps is learned.

– **SS-DTP** (Described Target and Previous time-steps), that introduces the year j_t and the month m_t of the target value to predict $y_{c,t}$, as well as the years $j_{t-1}, j_{t-2}, \ldots, j_{t-w}$ and the months $m_{t-1}, m_{t-2}, \ldots, m_{t-w}$ of the considered w previous observations.

It is noteworthy that, although SS-DT and SS-DTP explicitly represent the information about the year and the month associated with a given descriptive feature, the absolute value of a month does not properly represent the temporal cyclicity. In other words, December (12) 2018 may appear very distant to January (1) 2019, while it is actually temporarily close. To alleviate this issue, we resort to directional statistics that allow considering the *temporal position* of the target month, as well as that of the months historical data refer to (only in the case of SS-DTP). At this purpose, we use directional statistics that *envelope* the probability density function around the circumference of a unit circle representing the months of the year (see Fig. 2). More specifically, we compute the radial closeness between two months m_1 and m_2, represented as integer values in the interval $[1; 12]$, on the unit circumference as $2\pi - d_r(m_1, m_2)$, where:

$$d_r(m_1, m_2) = min\left(\frac{2\pi}{12} \cdot |m_1 - m_2|, 2\pi - \frac{2\pi}{12} \cdot |m_1 - m_2|\right) \quad (1)$$

is the radial distance between m_1 and m_2 on the acute angle (see Fig. 2 for an example of radial distance computed between February and May).

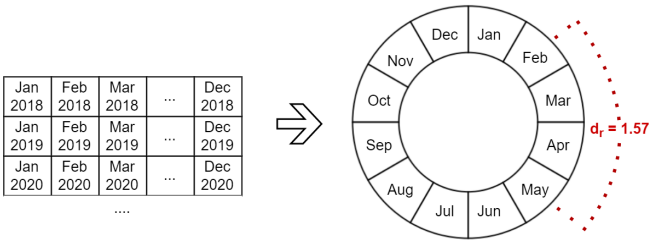

Fig. 2. Representation of the month of the year on the circumference of a unit circle. In the example, the radial distance between February and May is computed as $d(2,5) = min(2\pi/12 \cdot |2-5|, 2\pi - 2\pi/12 \cdot |2-5|) = min(1.57, 4.71) = 1.57$.

In our case, we compute the radial closeness between a given month in the descriptive attributes and the month of the target time-step to predict. Henceforth, the settings that exploit this radial closeness will be distinguished through a **C** (cyclical), appended to the name of the setting.

As regards the MS setting, we adopt the MIMO strategy to forecast 12 time-steps $y_{c,t}, \ldots, y_{c,t+11}$ for the consumer c at the same time. In this setting, we consider as input features the monthly energy consumption of the previous year (i.e., of the previous $w = 12$ months) and the year of the target time-step (see the right part of Fig. 1). Unlike the SS setting, MS does not need additional features to represent temporal relationships. Indeed, it is implicitly able to capture potential temporal dependencies, since the i-th feature always represents the i-th month of the year. On the other hand, while the recursive SS setting may be more suited when the training data is limited, MS preserves the dependencies also between the predicted values, and avoids the propagation of errors typical of the recursive SS strategy.

Note that, however, not all the regression methods can be easily extended to work in this setting. In our system, we adopt three different regressors, namely, Linear Regression, Regression Trees and Random Forests, also because of their ability to produce accurate models also when the available training data is poor.

Linear Regression methods aim to identify a linear model with coefficients $q = (q_1, q_2 \ldots, q_p)$, where p corresponds to the number of descriptive features plus 1 (the intercept), that minimizes the residual sum of squares between the observed target values in the training set, and the predictions provided by the linear approximation. For multi-step prediction, in our case, since we need to predict the consumption for the 12 subsequent months, identifying a predictive linear model corresponds to finding a matrix of coefficients $Q \in \mathbb{R}^{p \times 12}$ such that $\frac{1}{N} \sum_{i=1}^{N} ||u_i^\top Q - v_i^\top||_2^2$ is minimized, where $u_i \in \mathbb{R}^p$ is the vector of the descriptive features of the i-th training instance concatenated with a 1 (to take into account the intercept), N is the number of training instances, and $v_i \in \mathbb{R}^{12}$ is the vector of target values for the 12 subsequent months for the training instance u_i.

Learning methods for the construction of **Regression Trees** and ensemble thereof (e.g., **Random Forests**) are usually based on top-down induction pro-

cedures. Starting from the root node containing all the training instances, at each iteration, the *best* split, consisting of a descriptive feature and a threshold, is identified such that it well discriminates/separates the instances falling in the resulting children nodes. Leaf nodes of the tree store the actual predictions. The identification of the best split relies on some heuristics that, for regression tasks, are usually based on the reduction of the variance.

The extension of these approaches to solve multi-step tasks consists in storing multiple output values in the leaf nodes (12 in our case), and in a modified heuristics able to globally consider the contribution of the split towards the proper prediction of all the target values. Specifically, we adopt the *arithmetic mean of the variance reduction* computed over all the target time-steps.

3.2 Modeling the spatial autocorrelation

As mentioned in Sec. 1, taking into account the spatial autocorrelation in the construction of the predictive models may be beneficial in terms of accuracy, since spatially close consumers could exhibit a similar behavior, mainly due to similar climatic conditions. We evaluate the contribution coming from the adoption of two different spatial statistics [5]: the Local Indicator of Spatial Association (LISA) [1] and the Principal Coordinates of Neighbor Matrices (PCNM) [7].

According to [1], *i)* a LISA for a given observation must give an indication of the extent of significant spatial clustering of similar values around that observation, and *ii)* the sum of LISAs for all observations must be proportional to a global indicator of spatial association. In our case, given the set of n consumers, we first compute a neighborhood matrix $A \in \{0,1\}^{n \times n}$ as:

$$A[c_a, c_b] = \begin{cases} 1 & if\ dist(c_a, c_b) < maxDist \\ 0 & otherwise \end{cases} \tag{2}$$

where c_a and c_b are the a-th and the b-th consumers (with $1 \leq a \leq n$ and $1 \leq b \leq n$), $dist(c_a, c_b)$ is the geodesic distance between consumers, and $maxDist$ is a user-defined threshold on the maximum distance to consider the spatial autocorrelation phenomena among consumers as relevant. The matrix A is then normalized so that the sum of each row equals to 1[1], as follows:

$$A'[c_a, c_b] = \frac{1}{max(\sum_{i=1}^{n} A[c_a, c_i], 1)} A[c_a, c_b] \tag{3}$$

Using the matrix A', we can estimate the contribution of the neighborhood on each descriptive feature. Specifically, we first compute the z-score normalization for each descriptive feature x of each consumer c_a as:

$$x'_{c_a} = \frac{x_{c_a} - \mu_{x,c_a}}{\sigma_{x,c_a}}, \tag{4}$$

[1] Some rows in the normalized matrix can have a sum of 0, when the corresponding consumer has no other consumers falling in its neighborhood, according to $maxDist$.

where μ_{x,c_a} and σ_{x,c_a} are the average and the standard deviation of the descriptive variable x for the consumer c_a. Using the normalized value x'_{c_a}, we compute the spatial indicator I_{x,c_a} for the variable x of the consumer c_a as:

$$I_{x,c_a} = x'_{c_a} \cdot \sum_{i=1}^{n} (A'[c_a, c_i] \cdot x'_{c_i}) \tag{5}$$

The computed spatial indicators, one for each feature, can finally be added as additional descriptive features. Therefore, this solution leads to the introduction of w additional features, that represent the initial descriptive features influenced by the spatial closeness with other consumers.

A different approach to consider the spatial autocorrelation, as mentioned before, is represented by the PCNM. It allows us to extract additional, separate, spatial descriptive attributes, starting from the closeness among consumers. Its computation consists of the following main steps:

1. Compute a truncated squared distance matrix, as follows:

$$D^* = \begin{cases} dist(c_a, c_b)^2 & if\ dist(c_a, c_b) \leq maxDist \\ 4 \cdot maxDist & otherwise \end{cases} \tag{6}$$

   where $maxDist$ is a user-defined threshold.

2. Perform the Principal Coordinate Analysis (PCoA) [8] on D^*. This analysis consists in the diagonalization of Δ, where:

$$\Delta = -\frac{1}{2} \left(I - \frac{1 \cdot 1^\top}{n} \right) D^* \left(I - \frac{1 \cdot 1^\top}{n} \right) \tag{7}$$

with I be the identity matrix, and 1 be a vector of $1s$. After diagonalization, the principal coordinates are obtained by scaling each eigenvector of Δ by the square root of its correspondent eigenvalue. Note that the eigenvalues can be either positive or negative. Eigenvectors associated with high positive (resp., negative) eigenvalues represent a high positive (resp., negative) autocorrelation. Since we are interested in considering only positive spatial autocorrelation phenomena (i.e., spatially close consumers with similar behaviors, rather than spatially distant consumers with similar behaviors), only eigenvectors corresponding to positive eigenvalues are kept and used as spatial descriptors.

Henceforth, the settings that exploit the spatial dimension will be distinguished through **LISA** or **PCNM**, appended to their name. In Fig. 3, a graphical overview of all the proposed learning setting is provided, where the temporal, the spatial or both temporal and spatial dimensions are considered.

4 Experiments

In this section, we describe the considered real-world dataset and the experimental setting. Then, we show and discuss the obtained results.

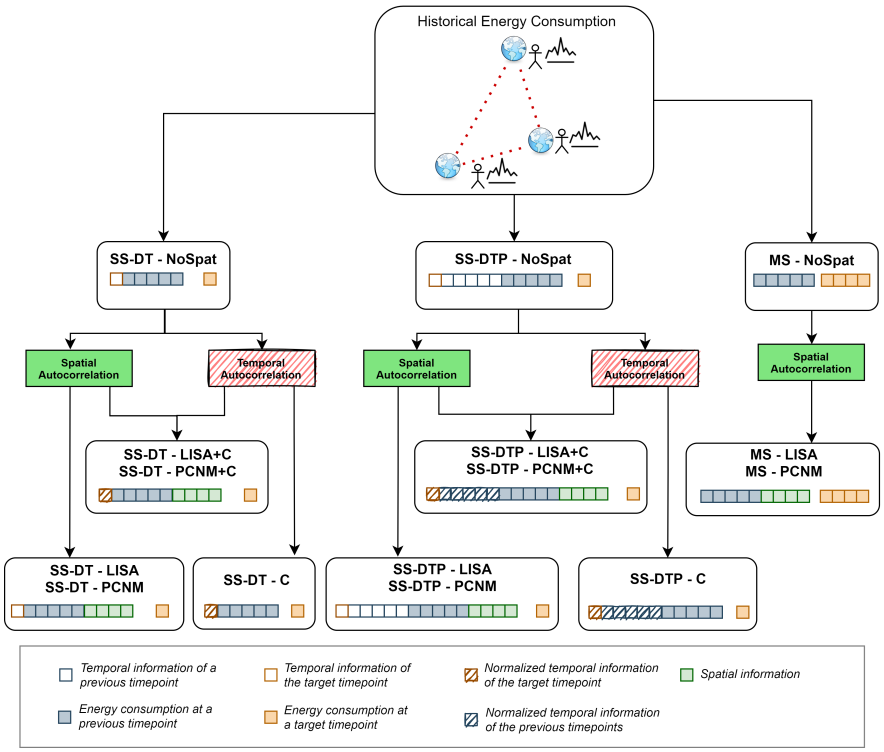

Fig. 3. A graphical overview of all the proposed learning setting is provided, where the temporal, the spatial or both temporal and spatial dimensions are considered.

4.1 Experimental Setting

We considered a dataset of an electrical network of a small city in the South of Italy consisting of 159 customers. Each customer is associated with the geographic coordinates (latitude and longitude) of the energy substation he/she is connected to in the network. The dataset consists of energy consumption data (in kWh) collected every month for a period of 10 years, i.e., from 2010 to 2019. Following a cross-validation setting for time-series, we iteratively consider each year from 2012 to 2019 as target year (see the quantitative information of the dataset in Tab. 1), with the goal of predicting the energy consumption for all the months of the target year, for all the customers of the network.

We performed the experiments with all the settings proposed in Fig. 3, to properly assess the contribution coming from the specific strategy adopted to take into account temporal and/or spatial autocorrelation phenomena. For LISA, we computed 12 indexes, one for each descriptive variable representing previous consumptions. For PCNM, we extracted 15 eigenvectors, following the experimental results reported in [5]. For both, the threshold $maxDist$ was set to 0.3 km, which is adequate in the context of a small city. As regressors, as introduced

Fold	Training period	Testing period	SS Training instances	MS Training instances
1	2010-2011	2012	1,908	159
2	2010-2012	2013	3,816	318
3	2010-2013	2014	5,724	477
4	2010-2014	2015	7,632	636
5	2010-2015	2016	9,540	795
6	2010-2016	2017	11,448	954
7	2010-2017	2018	13,356	1,113
8	2010-2018	2019	15,264	1,272

Table 1. Quantitative information of each fold of the considered dataset.

in Sec. 3.1, we considered Linear Regression (**LR**), Regression Trees (**RT**) and Random Forests (**RF**), available in *scikit-learn*. All the regressors were run with the default values for their parameters, except for the regression trees, for which we performed a grid search for the pruning parameter $ccp\_alpha \in \{0.2, 0.5, 1.0\}$. In Sec. 4.2, we report the best obtained results (i.e., with $ccp\_alpha = 1.0$).

As state-of-the-art competitor, we considered **GMAN** [20], a recently proposed neural network that is able to capture both spatial and temporal dimensions, through attention mechanisms, and of performing multi-step predictions. We adapted GMAN so that the temporal embedding encodes the month of each time-step, instead of the day and the hour, as in its original implementation. We also optimized its user-defined threshold ϵ on the spatial closeness, considering $\epsilon = 0.1$ (as suggested in [20]) and $\epsilon = 0.05$. In Sec. 4.2, we report only the best obtained results (i.e., with $\epsilon = 0.05$). Note that GMAN also performs a tuning phase on a validation set. Therefore, for this method, the results on the first fold are not available, since it requires data of an additional year as validation set.

As evaluation measure, we adopted the Relative Squared Error (RSE), which, contrary to other common measures like the RMSE, allows us to evaluate the predictive accuracy with respect to a simple predictor based on the average: a RSE close to 0.0 (resp. 1.0) means that the model has a perfect predictive accuracy (resp., equivalent to that of the simple average predictor), while a RSE over 1.0 means that the model is worse than the simple predictor. Formally, $RSE = \frac{\sum_t (r^t - \widetilde{r}^t)^2}{\sum_t (r^t - \overline{r})^2}$, where r^t and $\widetilde{r}^t$ are the true and the predicted values, respectively, for the t-th time-step, and $\overline{r}$ is the average value in the dataset.

4.2 Results and Discussion

In Tab. 2, we show the RSE result for each testing fold (target year), obtained by the considered regressors in the proposed settings, and by the competitor GMAN. We recall that the results of the first fold (2012) for GMAN are not available because it requires an additional year of data for its validation phase. Moreover, we do not report the results obtained in some settings of the LR (i.e., SS-DTP NoSpat, LISA and PCNM), since it was not able to fit a proper model (i.e., RSE>10) with the small amount of available training data for the first fold.

Looking at Tab. 2, we can make several observations. First, for the years 2012 and 2013, the RSE values appear quite high. This is due to the scarce availability

of training data for these folds (see Tab. 1). An exception is represented by the results obtained by MS, especially in the settings MS+PCNM and MS+LISA, that achieved good results also for these years. This may be due to the fact that the poor availability of historical data has been compensated by the captured dependencies among different time-steps and by the exploitation of the spatial information. Note that MS+PCNM appeared to be the setting that provided the best results overall for most of the years. Focusing on the regressors, the adoption of RF generally provided the best results in most of the settings, and when learned from the MS+PCNM setting, it led to the best absolute results. Note that, as emphasized in Sec. 3.1, learning methods for the induction of multi-step RTs and RF simultaneously optimize the construction of the model by considering all the time-steps. The capability of RF to reduce the variance in the predictions with respect to RT provided further improvements.

Looking at the results obtained by the considered state-of-the-art competitor GMAN, we can notice that, besides not being able to make predictions for the year 2012, the obtained RSE for the 2013 is very high, and quite close to the average baseline for the 2014. The RSE values become more acceptable for the subsequent years, but still higher than those achieved by the approaches proposed in this paper. These results prove that the approaches proposed in this paper to capture temporal and spatial autocorrelation phenomena are very effective with respect to those adopted by GMAN, and confirm the limitation of deep neural network architectures when the available training data is poor.

Overall, the strongest contribution appears to come from the MS setting. This observation is also clear from the average results shown in Fig. 4, where we can easily observe that the charts related to MS generally appear the lowest ones (i.e., with the lowest RSE), independently on the regressor. This confirms that the temporal dimension (and, especially, temporal autocorrelation phenomena) is fundamental for the prediction of the energy consumption in smart grids, and that capturing dependencies between different target time-steps provides higher advantages than explicitly representing the temporal information in the descriptive attributes, as done in the ST-DTP setting, and than adopting the radial temporal closeness (C). We further stress this aspect by observing the line charts in Fig. 5, where we plot the average RSE per month obtained by the best configurations according to Fig. 4 for each pair of setting (MS, SS-DT, SS-DTP) and regressor. From Fig. 5, we can observe that GMAN generally achieved an average high RSE, and that the MS setting led to more stable errors over the months of the year. This is due to its capability of capturing possible dependencies among the months of the year, and to avoid the propagation of errors introduced by recursive approaches. An interesting case is observable in the period April-May, where the highest prediction errors are made by almost all the approaches, probably due to the abrupt climatic changes that often happen in the South of Italy during such a period. On the other hand, the settings based on MS are able to provide accurate predictions also in these cases.

Table 2. Results in terms of RSE for each testing fold. The best result for each regressor (sub-table) and fold (column) is emphasized in bold, while the best result overall for each fold (column) is emphasized in bold with a gray background.

			2012	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
GMAN			-	9.640	0.840	0.650	0.370	0.280	0.270	0.359
Linear Regression	**SS-DT**	**NoSpat**	0.364	0.719	0.263	0.302	0.219	0.245	0.130	0.242
		LISA	0.366	0.726	0.264	0.305	0.221	0.251	0.133	0.243
		PCNM	0.392	0.757	0.279	0.301	0.225	0.247	0.131	0.252
		C	0.363	0.715	0.266	0.298	0.219	**0.241**	0.130	0.240
		LISA+C	0.364	0.720	0.266	0.302	0.221	0.247	0.133	0.241
		PCNM+C	0.391	0.752	0.281	0.298	0.224	0.243	0.130	0.249
	SS-DTP	**NoSpat**	-	**0.712**	0.260	0.289	0.214	0.249	**0.127**	**0.235**
		LISA	-	0.718	**0.258**	0.293	0.214	0.253	0.129	0.236
		PCNM	-	0.750	0.274	**0.288**	0.219	0.252	0.128	0.245
		C	0.402	**0.712**	0.261	0.290	0.214	0.250	**0.127**	0.236
		LISA+C	**0.359**	0.717	0.260	0.292	0.214	0.254	0.129	0.236
		PCNM+C	0.382	0.750	0.274	0.289	0.219	0.252	0.128	0.245
	MS	**NoSpat**	0.384	0.792	0.324	0.312	**0.205**	0.298	0.132	0.276
		LISA	0.417	0.862	0.332	0.330	0.235	0.333	0.136	0.284
		PCNM	0.394	0.826	0.350	0.319	0.214	0.302	0.134	0.282
Regression Trees	**SS-DT**	**NoSpat**	0.737	**0.805**	0.464	0.569	0.750	0.672	0.317	0.405
		LISA	0.774	1.866	0.540	0.430	0.619	0.915	0.295	0.448
		PCNM	0.820	0.885	0.428	0.412	0.432	0.688	0.302	0.374
		C	0.462	1.386	0.414	0.456	0.447	0.444	0.341	0.445
		LISA+C	0.437	1.429	0.527	0.472	0.469	0.684	0.351	0.390
		PCNM+C	0.492	1.634	0.446	0.521	0.380	0.766	0.392	0.366
	SS-DTP	**NoSpat**	0.624	1.361	0.500	0.631	0.480	0.377	0.336	0.475
		LISA	0.504	1.273	0.608	0.464	0.901	0.833	**0.274**	0.413
		PCNM	0.998	1.411	**0.397**	1.577	0.842	0.866	0.440	0.587
		C	0.851	1.072	0.556	0.689	0.694	0.562	0.332	0.504
		LISA+C	0.880	1.462	0.475	0.571	0.498	0.627	0.391	0.388
		PCNM+C	0.794	0.854	0.520	0.779	0.575	0.455	0.295	0.587
	MS	**NoSpat**	**0.364**	1.026	0.425	0.463	0.349	0.436	0.630	0.448
		LISA	0.443	1.096	0.732	0.502	0.307	**0.366**	0.323	0.435
		PCNM	0.460	0.984	0.454	**0.390**	**0.303**	0.467	0.337	**0.348**
Random Forests	**SS-DT**	**NoSpat**	0.300	0.893	0.336	0.307	0.197	0.572	0.132	0.251
		LISA	0.296	0.915	0.345	0.302	0.215	0.570	**0.127**	0.251
		PCNM	0.336	0.855	0.344	0.305	**0.188**	0.564	0.130	0.256
		C	0.320	0.912	0.326	0.310	0.211	0.553	0.135	0.244
		LISA+C	0.305	0.902	0.330	0.311	0.226	0.565	0.133	0.249
		PCNM+C	0.332	0.889	0.315	0.304	0.200	0.555	0.133	0.244
	SS-DTP	**NoSpat**	0.297	0.882	0.312	0.293	0.197	0.587	0.134	0.248
		LISA	0.302	0.876	0.333	0.306	0.217	0.595	0.131	0.251
		PCNM	0.331	0.869	0.324	0.297	0.196	0.573	0.133	0.248
		C	0.320	0.883	0.316	0.296	0.193	0.588	0.135	0.246
		LISA+C	0.303	0.904	0.327	0.294	0.208	0.592	0.128	0.247
		PCNM+C	0.325	0.855	0.320	0.298	0.199	0.568	0.132	0.249
	MS	**NoSpat**	0.262	0.578	0.254	0.277	0.195	**0.219**	0.148	0.234
		LISA	0.263	**0.520**	0.291	0.286	0.200	0.226	0.147	**0.229**
		PCNM	**0.259**	0.534	**0.253**	**0.270**	0.197	**0.219**	0.148	0.236

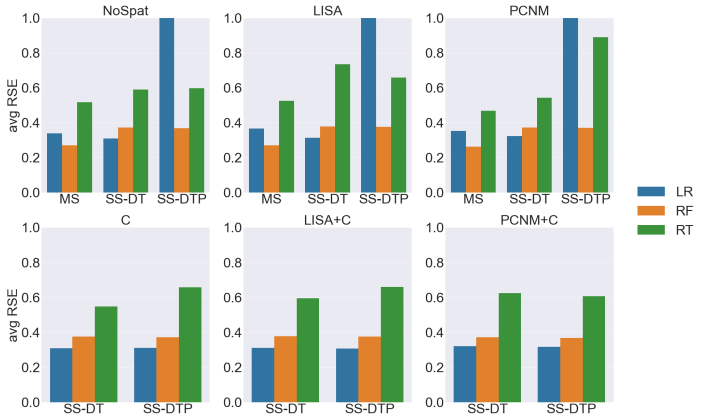

Fig. 4. Results in terms of average RSE. For readability, the results of LR in the upper part are graphically truncated to 1.0 (but they are actually around 1.5).

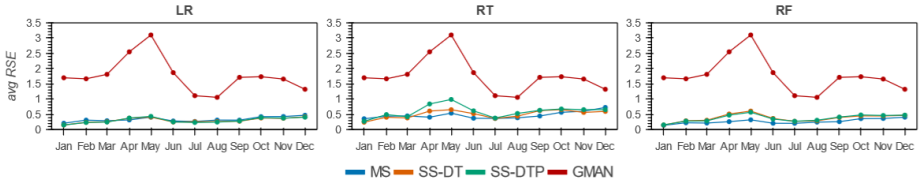

Fig. 5. RSE results averaged over the years for each month, obtained by the best configurations (see Fig. 4) for each pair of setting (MS, SS-DT, SS-DTP) and regressor.

5 Conclusion

In this paper, we proposed different approaches to take into account temporal and spatial autocorrelation phenomena while learning forecasting models for the prediction of the energy consumption in smart grids. For the temporal dimension, we investigated the contribution of the explicit representation of temporal information related to historical measurements, also through the temporal radial closeness, and that of predicting the value for multiple future time-steps simultaneously. For the spatial dimension, we investigated the contribution coming from the injection of LISA indexes and eigenvectors computed through the PCNM.

The experiments proved the overall superiority of models learned in the multi-step predictive setting, and the positive contribution coming from the PCNM, also when the available training data are scarce. The learned models also significantly outperformed the considered state-of-the-art competitor GMAN, which is based on a multi-attention neural network architecture.

For future work, we will consider the adoption of the proposed strategies for short-term predictions, in a nowcasting environment, and the integration of transfer learning techniques [11] to further improve the predictive accuracy when the available data related to a specific geographic area are poor.

References

1. Anselin, L.: Local indicators of spatial association — LISA. Geographical analysis **27**(2), 93–115 (1995)
2. Berriel, R.F., Lopes, A.T., Rodrigues, A., Varejao, F.M., Oliveira-Santos, T.: Monthly energy consumption forecast: A deep learning approach. In: 2017 Int. Joint Conference on Neural Networks (IJCNN). pp. 4283–4290. IEEE (2017)
3. Bontempi, G., Ben Taieb, S.: Conditionally dependent strategies for multiple-step-ahead prediction in local learning. Int. Journal of Forecasting **27**(3), 689–699 (2011)
4. Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control, 5th Edition. Wiley (2015)
5. Ceci, M., Corizzo, R., Fumarola, F., Malerba, D., Rashkovska, A.: Predictive modeling of PV energy production: How to set up the learning task for a better prediction? IEEE Transactions on Industrial Informatics **13**(3), 956–966 (2016)
6. De Assis Cabral, Joilson and Legey, et al. : Electricity consumption forecasting in Brazil: A spatial econometrics approach. Energy **126**, 124–131 (2017)
7. Dray, S., Legendre, P., Peres-Neto, P.R.: Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological modelling **196**(3-4), 483–493 (2006)
8. Gower, J.C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**(3-4), 325–338 (1966)
9. Lai, G., Chang, W.C., Yang, Y., Liu, H.: Modeling long-and short-term temporal patterns with deep neural networks. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 95–104 (2018)
10. Masum, S., Liu, Y., Chiverton, J.: Multi-step time series forecasting of electric load using machine learning models. In: International conference on artificial intelligence and soft computing. pp. 148–159. Springer (2018)
11. Mignone, P., Pio, G.: Positive unlabeled link prediction via transfer learning for gene network reconstruction. In: ISMIS 2018. pp. 13–23 (2018)
12. Ohtsuka, Y., Oga, T., Kakamu, K.: Forecasting electricity demand in Japan: A Bayesian spatial autoregressive ARMA approach. Computational Statistics & Data Analysis **54**(11), 2721–2735 (2010)
13. Serafino, F., Pio, G., Ceci, M.: Ensemble learning for multi-type classification in heterogeneous networks. IEEE Transactions on Knowledge and Data Engineering **30**(12), 2326–2339 (2018)
14. Shih, S.Y., Sun, F.K., Lee, H.y.: Temporal pattern attention for multivariate time series forecasting. Machine Learning **108**(8), 1421–1441 (2019)
15. Stojanova, D., Ceci, M., Appice, A., Džeroski, S.: Network regression with predictive clustering trees. In: ECML-PKDD 2011. pp. 333–348. Springer (2011)
16. Taieb, S.B., Bontempi, G., Atiya, A.F., Sorjamaa, A.: A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert systems with applications **39**(8), 7067–7083 (2012)
17. Tobler, W.R.: A computer movie simulating urban growth in the detroit region. Economic geography **46**(sup1), 234–240 (1970)
18. Williams, K.T., Gomez, J.D.: Predicting future monthly residential energy consumption using building characteristics and climate data: A statistical learning approach. Energy and Buildings **128**, 1–11 (2016)
19. Wu, Z., Pan, S., Long, G., Jiang, J., Zhang, C.: Graph wavenet for deep spatial-temporal graph modeling. arXiv preprint arXiv:1906.00121 (2019)
20. Zheng, C., Fan, X., Wang, C., Qi, J.: GMAN: A graph multi-attention network for traffic prediction. In: AAAI 2020. vol. 34 (01), pp. 1234–1241 (2020)