Rethinking weather station selection for electricity load forecasting using genetic algorithms

Santiago Moreno-Carbonell[a,*], Eugenio F. Sánchez-Úbeda[a], Antonio Muñoz San Roque[a]

[a] *Universidad Pontificia Comillas. ETSI ICAI. Instituto de Investigación Tecnológica*

Abstract

Demand forecasting is and has been for years a topic of great interest in the electricity sector, being the temperature one of its major drivers. Indeed, one of the challenges when modelling the load is to choose the right temperature, or set of temperatures, for a given load time series. However, minimum research efforts have been devoted to this topic. This paper reviews the most relevant methods that were applied during the GefCom2014 electricity demand forecasting competition and presents a new approach to weather station selection, based on Genetic Algorithms, which allows finding the best combination of temperatures for any demand forecasting model, and outperforms the results of those methods. In addition, our empirical results show that using different weights to combine the temperatures, optimized using the Broyden Fletcher Goldfarb Shanno algorithm, allows achieving significant error improvements.

Keywords: Electricity load forecasting, Weather station selection, Variable selection and combination, Genetic Algorithms, BFGS algorithm

1. Introduction

Energy forecasting has been of great interest from both academic and industry perspectives, especially, since the arrival of the firsts liberalized electricity markets. Among other relevant variables, electricity demand forecasting has

*Corresponding author
Email address: santiago.moreno@comillas.edu (Santiago Moreno-Carbonell)

been one of the most studied due to its importance for scheduling electricity generation, strategic bidding, making investment or regulatory decisions, and due to its importance as explanatory variable for other factors such as the spot price [1].

For years, all efforts have focused on point load forecasting and a wide amount of literature, including different approaches, have been published. However, the increasing market competition and renewable energy integration have caused a rapid growth of probabilistic load research. In [2], an exhaustive literature review about load forecasting and its evolution can be found. It is well known, as it can be seen in the vast majority of papers cited in [2], that the temperature is one of the main drivers of electricity demand, and their non-linear relationship has been widely studied. Because of that reason, the selection of the temperature or temperature combination to use for each problem is one of the firsts steps when modelling the load. Furthermore, electricity demand time series usually aggregate consumption from different geographical regions where there is typically available data from more than one weather station. For example, when the goal is to forecast the load of a given European country or a US state, the temperature of each city could be a candidate input variable for the model.

Weather station selection (WSS) methods aim at solving that problem. Given an electricity demand, their objective is to select the most relevant temperatures for that time series. As stated in [3], where authors proposed a systematic methodology for temperature selection, minimal research efforts have been devoted to this topic. In [3], the different approaches used in the GefCom2012 load forecasting competition are reviewed, raising two questions: (1) how many weather stations should be used and (2) which ones. Here, we will also focus on those two variable selection questions, and propose a new approach to WSS to allow answering both of them. However, we will also review two additional issues that are also relevant when applying any WSS method: (3) how to combine the different temperatures, and (4) for which model this combination is suitable. The third question is usually solved in the literature by averaging

the selected temperatures, but we will measure the impact of combining them using different weights. Regarding the fourth one, we will discuss the effect of using different models for the WSS method and the subsequent load forecasting model: is there an optimal temperature combination or, on the contrary, the goodness of the selection depends on the forecasting model we are going to use?

In summary, this paper proposes a new methodology for WSS for electricity load forecasting based on genetic algorithms, and discusses those four questions. First, the data from the GefCom2014 is described, as well as the benchmark model that has been used, including a brief description of the problem and the main approaches were used for WSS selection by the seven participant teams of the load forecasting competition presented in [4]. Next, the proposed GA approach for WSS will be presented, and its results, compared with the WSS method presented in [3] will be shown. Then, the potential improvements of using weights when combining the temperatures will be analysed. Finally, conclusions and future possible improvements will be presented.

2. Problem description

This section aims to describe the main features of the GefCom2014 forecasting competition, since it has been the source of data for the WSS method addressed in following sections. It also presents the benchmark model used in this paper, and briefly explains the different approaches of the top participants of the competition, focusing on their WSS methods.

2.1. GefCom2014 load forecasting competition and data description

The Global Energy Forecasting Competition of 2014 (GefCom2014) took place after the success of the previous edition of 2012 about hierarchical load forecasting, and dealt, in this case, with probabilistic forecasting in order to better capture the uncertainties of modern grids. All the data, rules, and main results are presented in [4]. In spite of the fact that it consisted of four forecasting tracks (load, price, wind and solar power), here we will only focus on the electricity load forecasting problem.

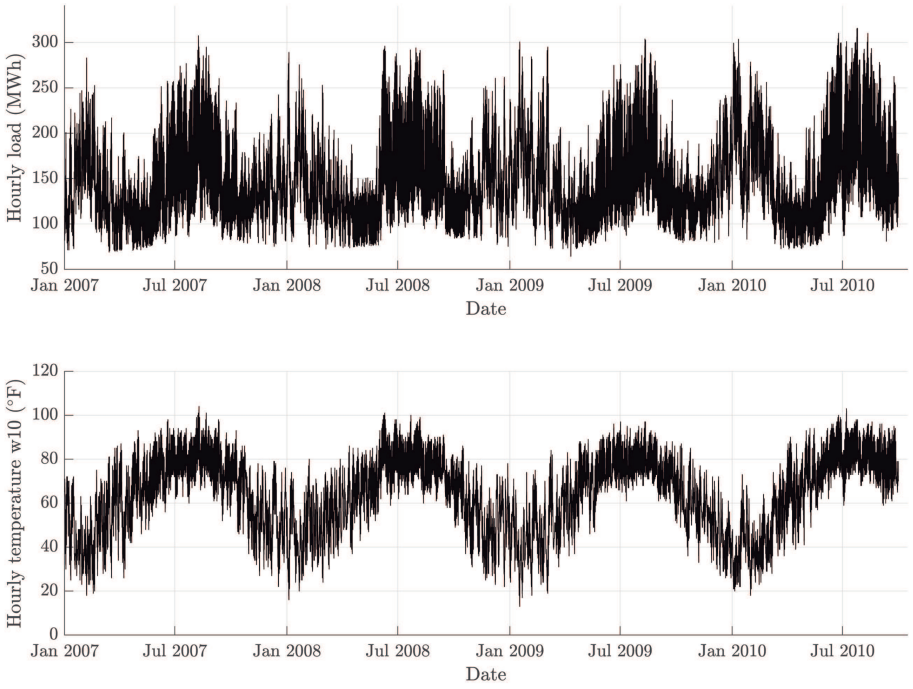

Figure 1: Load (*Top*) and temperature from the weather station w10 (*Bottom*) time series in GefCom2014 (years 2007-2010).

The information provided for that track consisted of 69 months of hourly electricity load data (a single demand time series), from 01/01/2005 to 30/09/2010 and 117 months of hourly weather data, from 01/01/2001 to 30/09/2010. Specifically, the weather data, consisted of 25 temperatures from different weather stations, of which there is no further geographical information. In addition, the list of US federal holidays could also be used. After this, during the competition, 15 additional months of data (weather and load) were provided. In this paper, in order to fairly compare our method with other methods proposed during the competition for WSS, only the first data release will be used.

In addition, let us describe the different data partitions that have been used in this paper. Using the same criteria of [5], where authors apply the same WSS method as the one presented in [3] with GefCome2014 data, the years

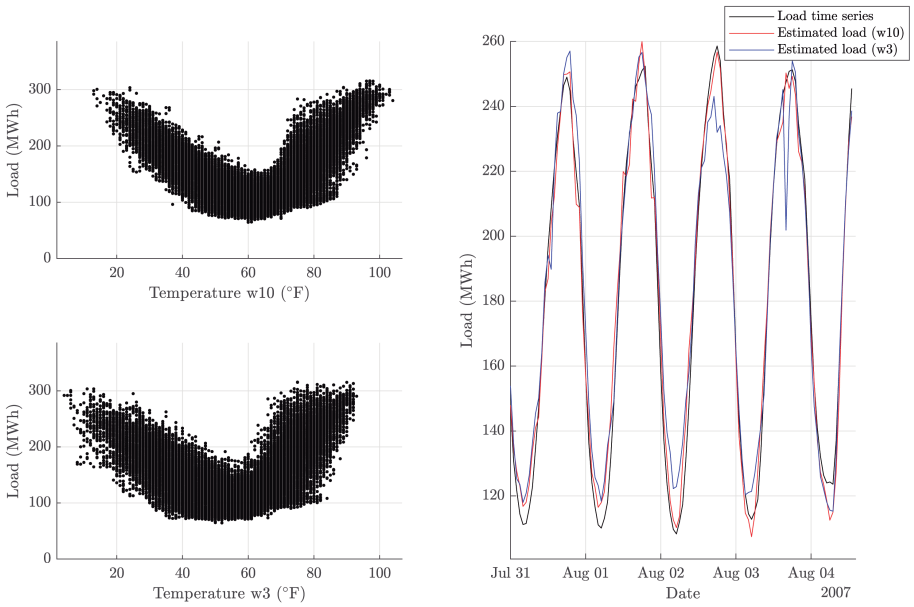

Figure 2: Nonlinear relationship between the load and the temperatures w10 (*Top-left*) and w3 (*Bottom-left*), and real and estimated load using these two temperatures and the benchmark model (*Right*).

2007 - 2009 will be used as training data (in-sample, 26304 points), and the year 2010 will be used as test set (out-of-sample, 6553 points). Figure 1 shows the load time series as well as one of the temperature time series (the one from the weather station w10) during both in-sample and out-of-sample periods.

As aforementioned, temperatures are one of the major drivers of electricity demand. These variables, together with calendar ones, such as the hour, the day of the week or the month are tipically the most used ones to forecast the load [6]. Modeling the well-known non-linear relationship between load and temperature can be determinant, and the accuracy of the model may depend largely on the selection of the right ones. As an example, Figure 2 shows this non-linear relationship using two different temperature time series. The selected temperatures (w10 and w3) are the best and the worst ones in terms of MAPE in the ranking presented in [5]. The differences between both responses can

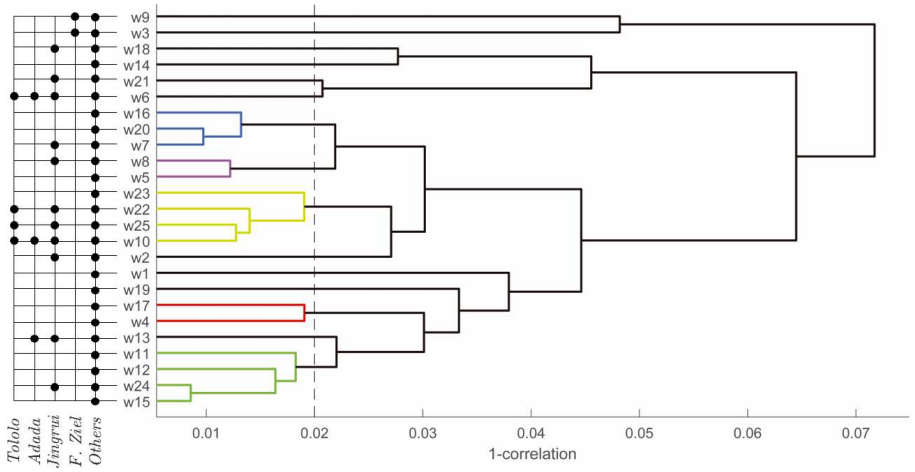

Figure 3: Dendrogram of the 25 weather stations of the GefCom2014 dataset. The dissimilarity threshold has been set to 0.02, resulting 15 clusters (resulting groups of temperatures are coloured). In the left part, the weather stations selected by the different competition teams are shown.

be seen in the figure, since the dispersion of the cloud of points when using the temperature w3 is noticeably greater and, therefore, the response of the demand is much less clear.

Furthermore, in order to quantify the effect of using one time series or the other, the benchmark demand model used during this paper, and described below in subsection 2.2, is fitted using both temperatures. The differences between the estimated load in both cases are illustrated in the week used as example in Figure 2 (*right*), and the error in the test set is measured in both cases. Using the temperature w10, the RMSE in the test set is 14.994°F, whereas using w3 is 22.514°F (more than 50% worse). This fact highlights the importance of a careful choice of temperatures before modelling the load, due to the huge impact in error that a poor choice can mean.

Finally, let us describe one of the the most common problems when selecting temperatures for forecasting electricity load, that is also present in this particular problem. When studying the demand of a given geographical region, some

of the candidate temperatures come usually from weather stations that are close to each other, leading to a great correlation between them. As it can be seen in the dendrogram of Figure 3, for example, considering a correlation threshold of 0.98, the temperature set could be reduced from 25 temperatures to only 15. This high correlation is usually a problem when selecting the proper variables when applying a WSS method, and some of the difficulties caused by this effect will be discussed below.

2.2. Benchmark model

Tao's Vanilla model (from now on, *Vanilla*) was first proposed in [7], and was used as benchmark during GefCom2012, being ranked in the top 25% among over 100 teams. It is a multiple linear regression model with the following effects:

- Main effects: linear trend, the first three order polynomials of the temperature (T, T^2 and T^3), and the categorical variables $Month$, $Weekday$ and, $Hour$.
- Cross effects: $Hour{*}Weekday$, $T{*}Hour$, $T^2{*}Hour$, $T^3{*}Hour$, $T{*}Month$, $T^2 * Month$ and $T^3 * Month$.

Besides being the benchmark during GefCom2012, this model was also used in [3] when presenting their WSS method, and it was also used by one of the participants of GefCom2014, (the *Jingrui* team, see [5]). Due to its simplicity and reproducibility, and in order to compare our results with theirs, this model will also be used as a benchmark in this paper.

2.3. WSS approaches in GefCom2014

In this section, the WSS approaches of the seven participant teams summarized in [4], including the top five entries of the competition, are discussed. In first place, the *Tololo* team [8], based their load forecasting model in generalised additive models (GAMs), creating a new virtual temperature by averaging four WS (6, 10, 22 and 25), selected using generalised cross validation (GCV). Their

results show that the resulting series leads to a significant improvement in GCV scores with respect to only using any of the stations separately.

Next, the *Adada* team [9] initially selected a set of six temperatures (6, 7, 10, 13, 22, and 25), also using GAMs and generalized cross validation. However, they did not obtain good results with that approach and decided to refine their WSS, calculating a weighted average using the exponentially weighted average algorithm (EWA, see [10]). Using this method, they finally chose three temperatures with equal weights: 6, 10 and 13. Noteworthy, this result can be analysed by observing the dendrogram presented in Figure 3. Their final three time series were included in their initial set of 6 temperatures selected with GCV. Two of the discarded temperatures (22 and 25) were highly correlated with station 10, that can be considered the representative of that cluster in their final election, which includes 3 temperatures that present fairly low correlations between them.

In third place, the team *Jingrui (Rain) Xie* [5] (from now on, *Jingrui*), used the WSS method described in [3]. Based on the *Vanilla* model, all the temperatures are initially ranked, measuring their in-sample MAPE. After that, the first n stations of the ranking are incrementally combined by averaging, with n between one and the total number of stations (25 in this case), the out-of-sample MAPE of each combination is calculated, and the best set is finally chosen. Using this method, the average of the first 11 temperatures of their ranking leads to the best result. Specifically, they finally selected the following stations: 2, 6, 7, 8, 10, 13, 18, 21, 22, 24, and 25. It is worth mentioning that this method, that adds in an incremental manner the variables taking into account the ranking of its individual in-sample MAPE, can lead to models with more variables than necessary. In this particular example, the first three temperatures of the ranking (stations 10, 22 and 25) have a correlation greater than 98% according to Figure 3, and this selection method would never allow removing useless variables, or to consider sets in which only one of this three temperatures would be chosen, for example. To illustrate the problems this may cause, the combination of three WS determined by their ranking is already

worse in this case than the set of three temperatures used by the *Adada* team (presents an out-of-sample RMSE 10.47% greater). Furthermore, only removing the w25 station from their final selection (the third one of their ranking, which is the one with highest correlation with w10), their out-of-sample error would improve 1.1%.

In fourth and fifth place, the teams *Oxmath* [11] and *E.S.Mangalova* [12] did not use any WSS method, and they simply used the mean of the 25 available temperatures. *Bidong Liu*, also listed in [4], and ranked eighth in the competition, used the same approach. Finally, *Ziel Florian* [13], which took second place in GefCom2014-E (an in-class extended version of GefCom2014 load forecasting competition, organized by Tao Hong and also open to external participants) and whose results were also presented in [4], chose the stations 3 and 9 because they gave the best in-sample fits to a cubic regression of the load against the temperature.

To conclude, Table 1 summarizes all these different approaches. The purpose of this table is to collect the different methods that have been used by each team, present the stations that have been selected in each case, and compare their in-sample and out-of-sample errors, using the benchmark model. In spite of the fact that all the presented methods have been tested here using the same criteria and forecasting model, it should be noted that the comparison of their errors is not completely fair. As aforementioned, the use of one load forecasting model or the other can determine the suitability of the different stations, and Table 1 is a clear example of that. The model structure or the use that each one makes of the different candidate explanatory variables (trend, calendar variables, temperatures...) can determine the best temperatures for each one. For that reason, we cannot conclude which set of temperatures is the best one in general terms, and our goal will be to find the best combination for one particular model: the *Vanilla* model (described in subsection 2.2).

In that sense, it can be seen that the WSS method presented by *Jingrui* team clearly outperforms the errors of the rest of models, but it is worth noting that their WSS method makes use of the *Vanilla* model to rank the variables to

Table 1: Summary of the different WSS methods applied during GefCom2014. The column *Stations* shows the WS that have been finally selected by each team. The two last columns show the RMSE, both in-sample and out-of-sample, using the *Vanilla* model

Team name	*Method*	*Stations*	RMSE (in-sample)	RMSE (out-of-sample)
Tololo [8]	Generalized cross validation (GCV)	6, 10, 22, 25	10.318	12.579
Adada [9]	Initially GCV. Refined using EWA	6, 10, 13	10.050	12.274
Jingrui (Rain) Xie [5]	Average of the n best stations from *Vanilla* model [3]	2, 6, 7, 8, 10, 13, 18, 21, 22, 24, 25	**9.523**	**11.772**
Ziel Florian [13]	Goodness of fit to a cubic regression	3, 9	18.552	20.920
Bidong Liu [13], *OxMath* [11], *E.S.Mangalova* [12]	Average of all the stations	1-25	10.462	12.700

be used. In this particular case, their results give us an idea of how much error reduction can be attributed to the use of the right WSS method (i.e. using the same model to choose temperatures and forecast the load). Comparing their results with those teams that did not use any WSS method, averaging all the temperatures, it can be seen that they achieved more than 7% error improvements both in and out-of-sample.

In a intermediate point, the teams *Tololo* and *Adada* present the second and

third best results, also outperforming the error of those teams that did not apply any WSS method, and averaged the 25 stations. Finally, the case of the team *Ziel Florian* is the clearest example of the huge differences that can mean the use of one forecasting model or the other when selecting temperatures. Attending to the goodness of the fit to a cubic regression between load and temperature, they selected the two best time series: the stations 3 and 9. However, these two temperatures are the worst two according to the in-sample MAPE ranking of [5], and also, the two with highest GCV scores in the ranking of [8]. As can be seen in Table 1, the use of the average of those two temperatures as input for the *Vanilla* model provides the worst in-sample and out-of-sample errors: in both cases more than 77% worse than the best option, and more than a 64% worse than not applying any WSS method. This fact highlights the dependency of the demand model when applying any WSS method, and gives us an idea of the huge difference in terms of error that using different WSS and forecasting models can mean.

3. Proposed GA approach

As it can be seen in the analysis of the previous section, the selection of how many temperatures and which ones must be used is not a trivial decision, and the goodness of the choice depend on the forecasting model that will be used afterwards.

For that reason, one of the main advantages of the method described in [3] is that, in spite of the fact that it is presented using the *Vanilla* model, it is a generic WSS method that allows finding a good temperature combination for any load forecasting model. Its simplicity and reproducibility are also great strengths of the method. However, its incremental nature, based on the error of each temperature separately, does not allow, for example, finding a combination without temperatures that are highly correlated, and can lead to an over-parametrized combination of temperatures.

In this section, we propose a new approach to WSS, based on genetic algo-

rithms (GA), that shares some features with the method presented in [3], but aims to solve its main drawbacks. Regarding their similarities, the GA approach allows to find the best temperature combination for any load forecasting model. As in [3], we will use the *Vanilla* model, which will be the fitness function of the GA, due to its simplicity and relative computational simplicity. Regarding how to combine the temperatures, we will also follow a similar approach to [3], and will be combined by averaging. However there are important differences between both methods. The main advantage of our GA approach is that it does not add the temperatures incrementally, what allows testing, for any number of temperatures (K) the optimum combination independently of the selected ones in the execution with $K-1$, thus avoiding the problem of selecting sets with high correlated variables.

GA are heuristic optimization methods based on natural evolution, and are explained in detail for example in [14] and [15]. In summary, a first population of individuals (representing different candidate solutions) is initially created, then their individuals evolve iteratively, generation to generation, achieving better and better results attending to an objective (*fitness*) function. The fitness of each individual of the population is calculated in each iteration, being the better candidates more likely to stay in the following generation. The creation of the next generation from a previous one depends on two operators: *crossover* and *mutation*. Furthermore, if the elitist selection is present, the best candidates of the population, the *elite*, are kept unchanged to the next generation. The nature of this kind of algorithms make them suitable for variable selection, and several examples of its use can be seen in [16], [17] or [18]. However, we have not found any GA approach in the literature regarding WSS methods. Here, the proposed WSS-GA method uses *Vanilla* to compute the fitness function, and has specific initialization, crossover and mutation functions that will be explained below. It has been implemented using the Global Optimization Toolbox of MATLAB.

Firstly, let us describe the genetic representation and the parameters that have been used. Following the approach described in [16], each individual is encoded as a set of N 0's and 1's, being N the total number of input candidate

variables (i.e. the 25 weather stations time series when using the data from GefCom2014), and being K the number of selected variables by an individual (i.e. the number of ones). For example, in a variable selection problem with $N = 8$ candidates, the *chromosome* 01001001 would represent an individual with $K = 3$, that will use the variables 2, 5 and 8 in the modelling process, omitting the other five ones.

Concerning the fitness function, given a particular individual, the average of its selected temperatures is computed and then, using this new virtual temperature as input, the *Vanilla* model is fitted to the in-sample data. Finally, its in-sample RMSE is measured, which is used later as fitness score.

Regarding the composition of each generation of the population, 5% of the population forms the *elite*, the best performing individuals of each generation that keep unmodified in the next one. Then, the 80% of the remaining individuals will generate *crossover* children, and the rest, will generate *mutation* children. Before describing the initialization, mutation and crossover functions, let us explain the execution process that we followed, since it has conditioned the design of the different operators.

As aforementioned, the goal of our GA approach is to find how many and which temperatures must be used. In order to answer that two questions we have followed a similar approach to the one used by the *Jingrui* team in [5], where the error of the combinations of weather stations for any value of K is presented (from only one, to averaging the 25 variables). Here, the proposed GA will be executed N times, keeping the number of selected variables (K) constant in each execution. The initialization, mutation and crossover functions have been specially designed to fulfill that requirement: instead of adding constraints to the GA, all the individuals are randomly initialized to feasible ones (i.e. individuals with K selected variables), and the mutation and crossover functions do not allow the appearance of infeasible chromosomes. Figure 4 shows an illustrative example of the initialization, crossover and mutation processes, where there are $N = 8$ candidate variables and $K = 3$.

Firstly, regarding the *initialization* process, as many individuals as the popu-

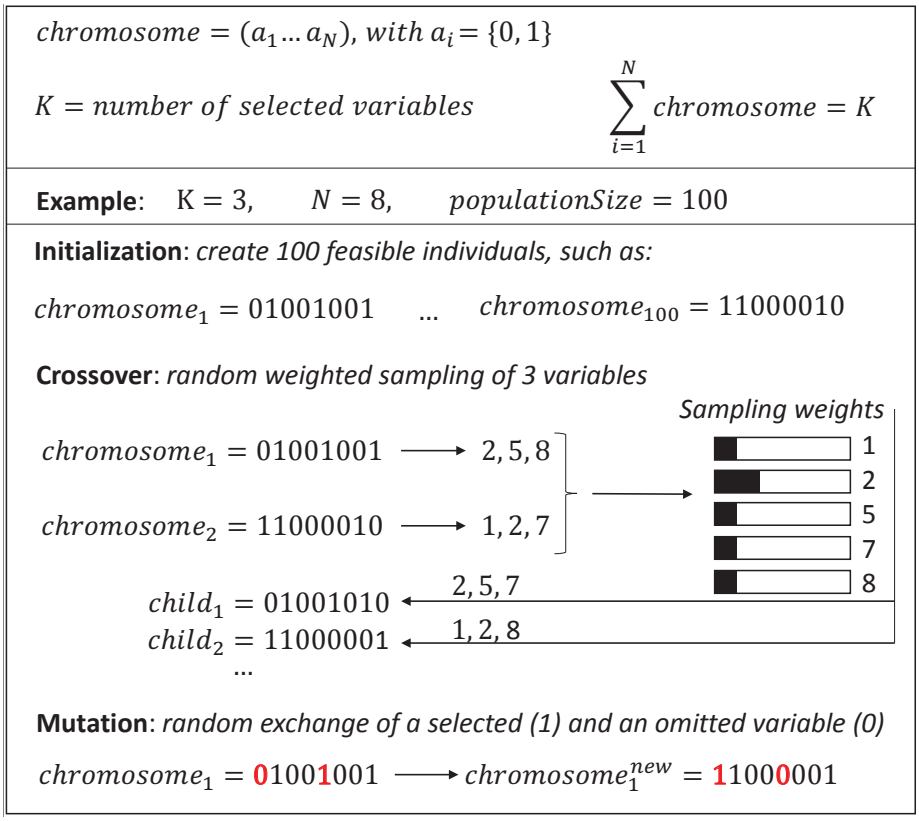

Figure 4: Illustrative example of the initialization, crossover and mutation functions of the proposed GA. Variable selection problem with $N = 8$ candidate variables and $K = 3$ selected variables is shown. Using these three functions, the feasibility of each individual (i.e. $K = 3$) is granted during the GA while the problem keeps unconstrained.

lation size indicates are generated. It can be seen that only individuals with the right number of selected variables (K) are always created. Furthermore, as the population size can be greater than the total number of possible variable combinations for some values of K, all the feasible different individuals are forced to be generated in that cases. On the other hand, if the population size does not cover all the search space, all the randomly generated individuals are then forced to be different. In the example of Figure 4, as there are only ${}^{8}C_{3} = 56$

possible subsets of three different variables among the 8 candidates (full search space), creating 100 individuals during the initialization process would force the generation of that 56 different individuals and some redundant ones. It should be noted that doing that would be equivalent to carry out the most straightforward solution to this problem: examine all the possible combinations of the variables exhaustively, a 'brute force' solution only feasible when the number of combinations is small as in this case. For example, in the GefCom2014 variable selection problem, we start with a search space of 25 points when $K = 1$, but in the worst case scenario, given by the combinations of 12 and 13 different variables ($^{25}C_{12}$ and $^{25}C_{13}$), there are more than $5 \cdot 10^6$ possible combinations. In fact, to test all the possible temperature combinations exhaustively in the GefCom2104 problem (for any value of K), more than $33 \cdot 10^6$ calls to the fitness function would be required ($\sum_{i=1}^{25} {}^{25}C_i$).

Secondly, regarding the crossover function, the aim will be to generate a crossover child for two given individuals (*parents*), keeping constant their number of selected features, and having into account the importance of each one. To do that, all their candidate variables are initially listed and, according to their occurrence, their weights are calculated in order to carry out a weighted random sampling, selecting then K of that variables. In the example of Figure 4, both chromosomes contain a total of 5 different candidate variables, appearing twice the variable number two, so it will have twice as much weight than the other four in the subsequent random sampling, and therefore will be twice as likely to appear as a selected variable in the crossover child.

Finally, the *mutation* function follows a quite simple approach, it just exchanges a zero by a one of the *chromosome*, both randomly selected. In the example, the initially omitted variable number one is selected by the mutation child, whereas the variable five stops being one of the selected ones. It should be noted that these two functions, together with the initialization one, allow following a double objective: ensuring that all the individuals have a constant K and favouring the presence of the most important variables for each generation of the GA.

4. Results

In order to follow a similar approach to [5] and build the error curve for any value of K, the GA is executed 25 times independently, obtaining the best temperature combination at each step. Before each execution of the GA, the initial population is created, generating a set of individuals with the correct value of K. Regarding the population size, several tests have been done, varying from 200 to 10000 individuals, and a value of 1000 have been finally chosen for all the executions. It should be noted that for those values of K that lead to a number of possible combinations less than a thousand (i.e. ${}^{25}C_1 = 25$, ${}^{25}C_2 = 300$, and ${}^{25}C_{25} = 1$) all the search space has been exhaustively explored and the best temperature combination has been definitely found. It is also remarkable that the tests that have been carried out by randomly varying the initial population (initial search point), with population sizes of 1000 and 10000, have provided the same results in terms of selected variables and therefore, error.

Figure 5 shows the order in which the different weather stations have been selected by the GA with different values of K, compared with the ones used in [5] after ranking the variables according to their in-sample error. It can be seen that they are noticeably different. First, as aforementioned, the combination of *Jingrui* begins with the weather stations 10, 22, and 25, in spite of the fact that they are highly correlated. The GA begins adding the variable 25, but when $K = 2$ that variable is removed and the average of variables 6 and 10 is selected, and, as will be shown later, that combination has already less error than the combination chosen by *Jingrui* for that value of K. From that point, the weather station w6 is systematically selected by the GA in every execution.

Regarding the behaviour of the GA in that entry sequence of variables, Figure 5 shows that we cannot make an unique ranking of the best temperatures attending to their order of selection, but there are only three points that break that idea of incremental sequence. To begin with, the variable 25 is selected in first place, and it is immediately removed when $K = 2$. After that, it will not be selected since the combination of 11 weather stations. The second breaking

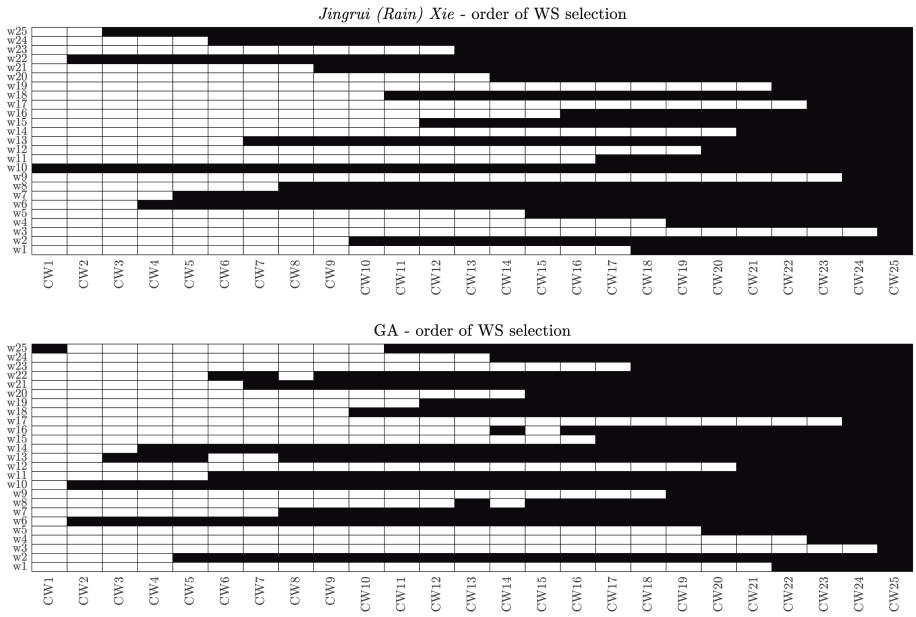

Figure 5: Order of variable selection in both the WSS method of [5] and the proposed GA. Each column represents the set of selected temperatures for a given value of K, and each row a weather station (1-25).

point appears between the combination CW6 and CW8, when the variable 22 seems to replace the variable 13, being omitted afterwards in the combination of $K = 8$, when the station 13 is selected again. The last case, similar to the second one, begins with $K = 14$, when the weather station 16 is selected at the expense of the variable 8, being removed in the next execution of the GA, and selected again when $K = 16$. This behaviour may be due to the independence of one execution of the GA and the following, the randomness of the initial search point and, above all, the high correlation between the temperatures that can lead to several similar solutions even combining different temperatures. For example, it is known the high correlation of the variables 10 and 25 that are exchanged when $K = 2$. It is also remarkable that, for the first two values of K, the search space have been exhaustively covered, so that these two combinations are the best ones in terms of in-sample RMSE, thus replacing w25 by w25 in

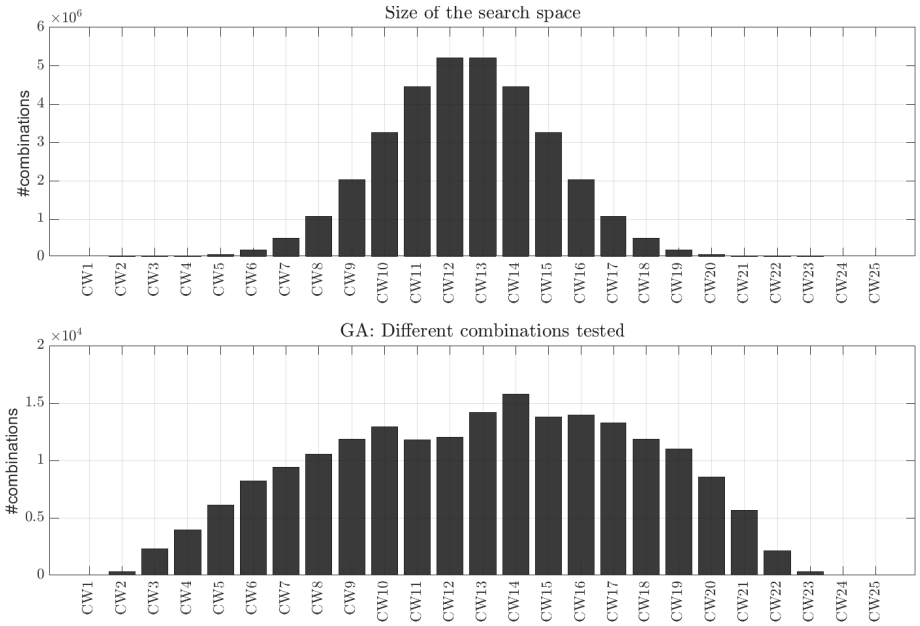

Figure 6: Size of the search space for each value of K (*Top*) and actual number of combinations tested by the GA at each iteration (*Bottom*).

CW2 is correct.

On the other hand, regarding the size of the search space, given by the total number of different temperature combinations that can be explored at each iteration, Figure 6 shows its actual size for each value of K ($^{25}C_K$), and the actual number of options that has been covered by the GA search in each execution. It can be seen that, as expected, the bigger the search space is, the greater the number of combinations required by the GA to find the optimum solution is. However, it should be noted that, far form covering exhaustively that search space, the algorithm is obtaining a good solution exploring only a small part of it skilfully. For example, the most extreme case is given for $K = 12$, where there are more than $5 \cdot 10^6$ ($^{25}C_{12}$) different possible combinations, and only 0.23% of them has been required by the GA to find an optimum solution. Considering the whole set of executions, only 0.34% of the full search space has been covered.

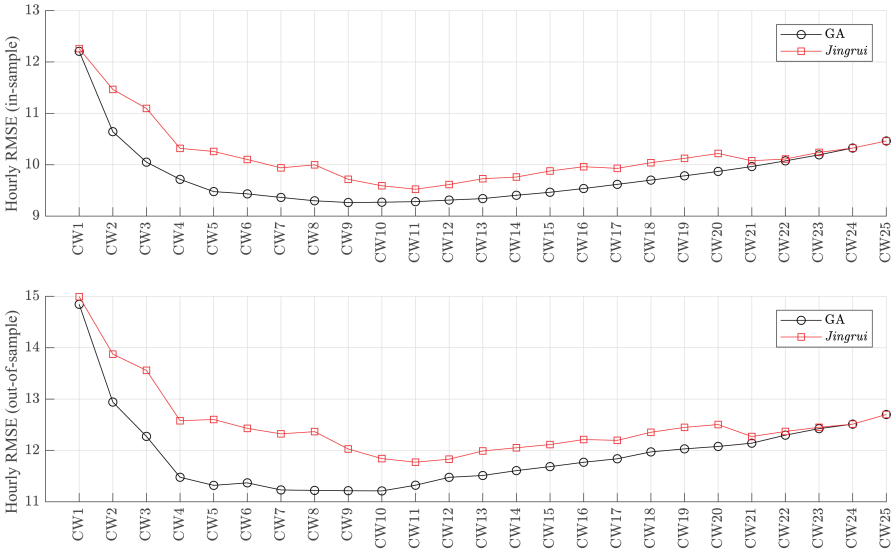

Figure 7: Error curves of the different temperature combinations, using the method of [5], and the proposed GA. The in-sample error curves (*Top*) and the out-of-sample error curves (*Bottom*) of both methods are shown.

Finally, Figure 7 shows the in and out-of-sample error curves obtained for each value of K, using the set of temperatures selected by the GA at each step, and also the ones formed using the ranking of the *Jingrui* team. For more detail, and in order to easily compare the results of both WSS methods for each value of K, Table 2 shows the sequence of selected variables of both methods, and their in-sample and out-of-sample errors.

From the very first points of the curves, the combination chosen by the GA outperforms the one selected by the other method. In this case, the fact of adding the station number 6 when $K = 2$, instead of continue adding the stations 10 and 22, very correlated with the first one, turns to be determinant in terms of error. It is remarkable that the first point of the curves of both methods are not the same, but very similar, in spite of the fact that in both cases it should be simply the best station in terms of error in the in-sample set: the GA selects the weather station 25 and the *Jingrui* team selected the variable

Table 2: Summary of the sequence of weather stations selected for each value of K by both the GA and *Jingrui* team methods. The two last columns show the in-sample and out-of-sample errors, using the *Vanilla* model. The values with * represent the lowest error for each method. The bold value represent the lowest out-of-sample error reached.

K	Added stations		Removed stations	RMSE (in-sample)		RMSE (out-of-sample)	
	Jingrui	GA	GA	*Jingrui*	GA	*Jingrui*	GA
1	10	25	-	12.260	12.207	14.994	14.844
2	22	6,10	25	11.465	10.643	13.875	12.942
3	25	13	-	11.095	10.050	13.559	12.274
4	6	14	-	10.318	9.715	12.579	11.478
5	7	2	-	10.257	9.477	12.603	11.321
6	24	11,22	13	10.101	9.432	12.430	11.369
7	13	21	-	9.939	9.363	12.323	11.231
8	8	7,13	22	9.998	9.298	12.367	11.222
9	21	22	-	9.716	9.264	12.029	11.217
10	2	18	-	9.590	9.270	11.842	**11.212***
11	18	25	-	9.523	9.281	11.772*	11.324
12	15	19	-	9.614	9.312	11.830	11.477
13	23	8	-	9.727	9.340	11.990	11.513
14	20	16,24	8	9.760	9.405	14.994	11.608
15	5	8,20	16	9.878	9.464	12.115	11.686
16	16	16	-	9.961	9.536	12.213	11.770
17	11	15	-	9.929	9.617	12.196	11.838
18	1	23	-	10.039	9.700	12.354	11.972
19	4	9	-	10.124	9.784	12.450	12.031
20	12	5	-	10.218	9.869	12.505	12.078
21	14	12	-	10.077	9.964	12.271	12.142
22	19	1	-	10.111	10.075	12.370	12.298
23	17	4	-	10.239	10.192	12.450	12.425
24	9	17	-	10.324	10.324	12.512	12.512
25	3	3	-	10.462	10.462	12.700	12.700

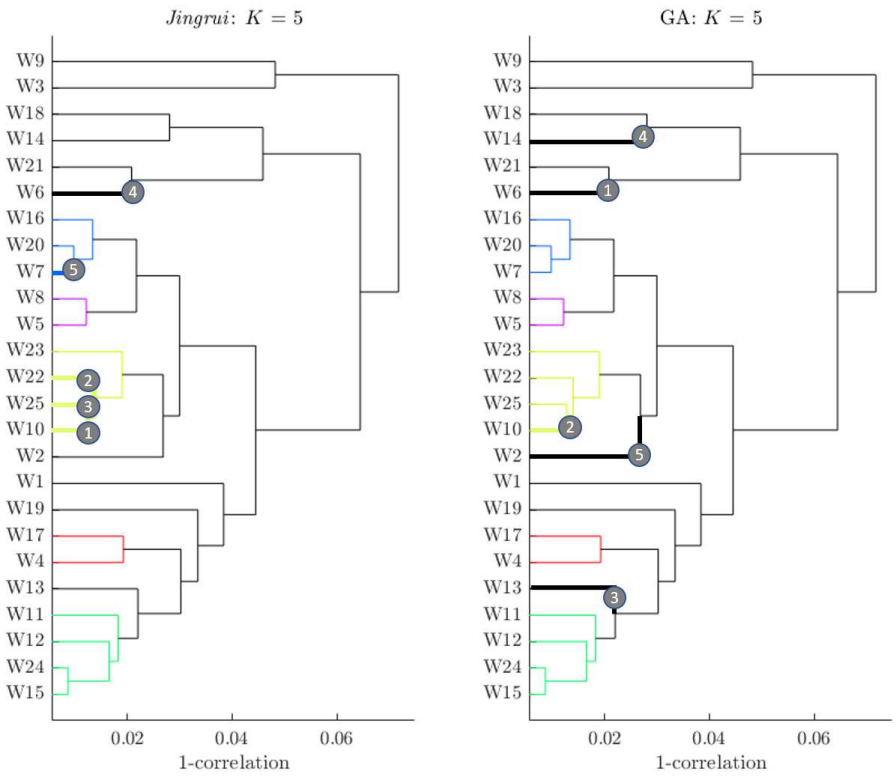

Figure 8: Weather stations selected for $K = 5$ by the two WSS methods: *Jingrui* (*Left*) and the proposed GA (*Right*). The coloured clusters of the in-sample dendrograms correspond to those formed using a correlation threshold of 0.98.

number 10. This fact is due to the difference in the error metric used in each case: in spite of the fact that these two stations are the best ones using RMSE or MAPE, the station 25 is the one with the lowest RMSE (12.207 against 12.260), whereas the station 10 is the one with the lowest MAPE (9.100 against 9.280), both in the in-sample set.

Comparing the evolution of the error in both methods, it can be seen that the curves from the GA present a much deeper decrease in error, mainly from $K = 1$ to $K = 5$, while the *Jingrui* method, not only decreases in a slower way, but also, it does not reach the errors of the GA. As can be seen in Figure 8,

Table 3: Summary of the in-sample and out-of-sample errors obtained with the *Vanilla* model, using the temperature combinations selected by *Jingrui* team and the GA. The last two columns show the results of the Diebold Mariano statistical test, which compares the predictive accuracy of the forecasts of both methods.

	Jingrui ($K = 11$)	GA ($K = 10$)	Error improvement	DM test	
	RMSE	RMSE	$\Delta RMSE(\%)$	*Statistics*	*p-value*
In-sample	9.523	9.270	2.657	12.228	0
Out-of-sample	11.772	11.212	4.757	13.229	0

where the combinations of both methods when $K = 5$ are shown over the dendrogram of Figure 3, the reason of these differences in error is clear. Whereas the WSS incremental method of *Jingrui* is adding correlated variables first, the GA takes advantage adding other relevant time series. The cluster formed by temperatures 10, 22, and 25, is represented in the GA set by w10, that enters when $K = 2$, being the other 4 variables determinant for reducing the error.

This last figure evidence one of the main problems of the method presented in [3]: in spite of averaging all the different temperatures with equal weights, the fact of adding temperatures in a sequential way based on an individual error ranking can lead to suboptimal solutions. Furthermore, we can be assigning greater weight to a given temperature or geographical area if there are several very similar temperatures in the candidate set. If the temperatures of this problem came from a country with five different areas equally relevant to the demand, and there were more than one available temperature from one of them (i.e. if all the temperatures the cluster formed by the stations 10, 22, 23, and 25 came from a given zone), using the individual error ranking as selection criteria would cause that the area would have three times the weight of the others, and even, if we would select the combination with $K = 5$, we would ignore the other two.

To conclude, and attending to Table 2, using the GA approach the minimum

error in the out-of-sample set is obtained for the combination of $K = 10$ stations, so that would be our selected set of temperatures. Specifically, the selected weather stations would be: 2, 6, 7, 10, 11, 13, 14, 18, 21, and 22. Table 3 summarizes the errors of the compared WSS methods, and includes the results of the Diebold-Mariano statistical test [19], in order to measure the significance of the predictive accuracy improvement obtained using the temperatures selected by the GA. It can be seen that we obtain significant improvements, in and out-of-sample, over the error of the method of the *Jingrui* team, with a time series less. Finally, it should also be noted that the out-of-sample error curves provided by the GA are quite flat from the combination of $K = 7$ to $K = 10$ so we could choose a set with even less temperatures, keeping an improvement of 4.6% over the other method.

5. Combining temperatures with different weights

This last section aims at measuring the impact of combining the selected temperatures using different weights instead of simply averaging. In spite of the fact that the most common approach in GefCom2014 was to average the selected time series, several examples of weighted combinations can be found in the literature. For example in [20], the weighted average by economic factors and load level per region are tested, concluding that the average was superior in that case. Here, the error improvements achieved optimizing the weights of the selected temperatures will be measured.

Since in GefCom2014 no geographical or economical information about the different weather stations was provided, approaches like those described in [20] can not be applied in this case. Our goal will be to find the optimum weights to the set of K temperatures that has been selected using the GA, minimizing the error when using the same load forecasting model. The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [21] is a quasi-Newton method for unconstrained non-linear optimizations problems. This method, and specially, its simplified limited-memory version (L-BFGS) [22] has been widely used in the

Table 4: Summary of the in-sample and out-of-sample errors obtained with the *Vanilla* model, using the temperature combinations selected by *Jingrui* team and the GA, and combined by averaging or using the optimized BFGS weights. The last two columns show the results of the Diebold Mariano statistical test, which compares the predictive accuracy of the forecasts with and without BFGS.

Method		Average	BFGS weights	Error improvement	DM test	
		RMSE	RMSE	$\Delta RMSE(\%)$	*Statistics*	*p-value*
Jingrui	In-sample	9.523	9.380	1.527	10.386	0
	Out-of-sample	11.772	11.541	2.001	6.618	3.64e-11
GA	In-sample	9.270	9.225	0.492	6.469	9.89e-11
	Out-of-sample	11.212	11.127	0.769	4.876	1.08e-6

literature to optimize parameters in diverse machine learning problems, such as [23] or [24]. Using this algorithm, the Hessian matrix of second derivatives is not computed, but it is approximated using updates specified by gradient evaluations. In our case, these updates will be given by different evaluations of the *Vanilla* model, varying the weighted average of the previously selected temperatures. Therefore, here the variables to optimize will be the set of weights of the K selected temperatures, whose module will be set to one before each iteration.

Following the same methodology of previous sections, the in-sample data will be used to optimize the weights and the error will be measured both in and out-of-sample. To illustrate our results, Table 4 shows the errors obtained before and after optimizing with BFGS, and Figure 9 shows the weights obtained for each temperature using the temperature sets selected by both WSS methods (*Jingrui* team and GA). Observing Table 4 and Figure 9, several conclusions can be drawn:

- The average is not the best option: optimizing the weights provides significant error improvements both in and out of sample, and for both WSS methods. In this case, the out-of-sample error improvements achieved

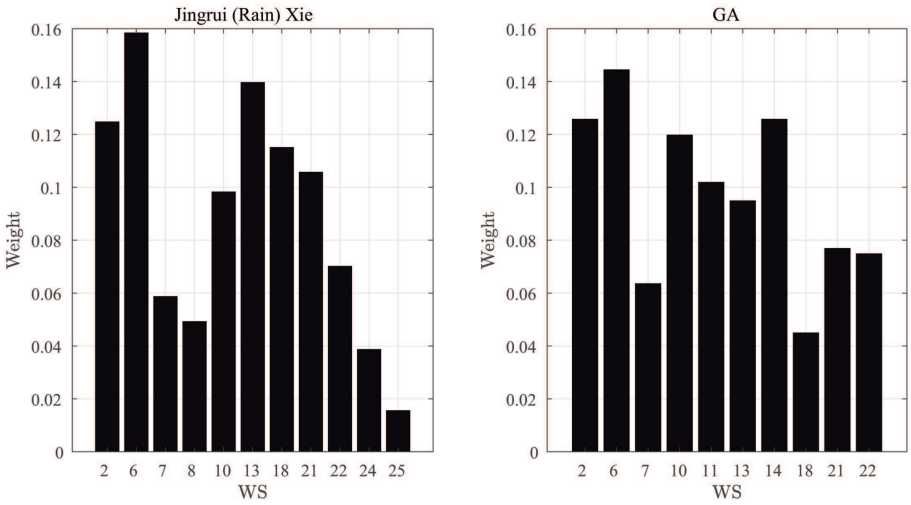

Figure 9: Weights obtained using BFGS to minimize the error using *Vanilla* model, and the temperature sets determined by the method of *Jingrui* team [5] and the proposed GA.

against the average vary between 0,77% and 2%.

- The temperature combination of *Jingrui* team always provides greater errors than the GA, even when comparing its results using optimized weights against the GA using the average.
- GA combination seems to have less room for improvement: the error reduction obtained with *Jingrui* temperatures is nearly three times greater than the ones obtained optimizing the GA ones (both in and out of sample). As can be seen in Figure 9, BFGS is reducing the effect of some temperatures, such as w25, that has a weight of 1.5%. Weights of GA temperatures are more equilibrated: w18 is the one with lowest weight (4,5%), but removing them from the set increases the final error.
- BFGS can decrease the effect of temperatures that should not be in the final set but, obviously, can not add not considered ones.

In summary, this section has proven that combining the temperatures using different weights can provide better results than simply averaging. The fact of

not having geographical or economical information about the problem does not prevent using different weights: in this case, using BFGS improves our final results, both in and out-of-sample (they are not over-fitted). Finally the fact of having less room for improvement in the GA case than in the other method, and having obtained a set of balanced weights, highlights once again the goodness of the previous temperature selection of the GA.

6. Conclusions

Temperature is one of the most important drivers of electricity load, and selecting the right station or combination of stations is crucial in terms of error. Furthermore, there is not a unique set of optimum temperatures to be chosen: the structure of the forecasting model, or the use that it makes of the different explanatory variables (trend, calendar variables, temperatures...) can determine the best stations for each one. In this paper we analyse the different weather selection methods (WSS) used during the GefCom2014 load forecasting competition, and propose a new approach, based on genetic algorithms (GA), that can be used with any load forecasting model, presents an easily reproducible implementation, and outperforms the results of the other presented methods. Furthermore, we analyse the effect of optimizing the weights of the selected temperatures using BFGS, concluding that the average is not always the best option, proving the goodness of the variable selection made by the GA, and reducing the error achieved by the GA approach and other alternatives from the GefCom2014 competition. Finally, this paper deals with the problem of creating a virtual temperature time series to be used as input to a load forecasting model. As next steps, the possibility of selecting a sub-set of temperatures to be used by the forecasting model will be tested, in order to consider the different responses of the demand to several temperatures or the importance of one or the other weather station depending on the time period.

References

[1] R. Weron, Electricity price forecasting: A review of the state-of-the-art with a look into the future, International Journal of Forecasting 30 (4) (2014) 1030–1081. doi:10.1016/j.ijforecast.2014.08.008.

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

[3] T. Hong, P. Wang, L. White, Weather station selection for electric load forecasting, International Journal of Forecasting 31 (2) (2015) 286–295. doi:10.1016/j.ijforecast.2014.07.001.

[4] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, R. J. Hyndman, Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond, International Journal of Forecasting 32 (3) (2016) 896–913. doi:10.1016/j.ijforecast.2016.02.001.

[5] J. Xie, T. Hong, GEFCom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation, International Journal of Forecasting 32 (3) (2016) 1012–1016. doi:10.1016/j.ijforecast.2015.11.005.

[6] A. Muñoz, E. F. Sánchez-Úbeda, A. Cruz, J. Marín, Short-term forecasting in power systems: a guided tour, in: Handbook of Power Systems II, Springer, 2010, pp. 129–160. doi:10.1007/978-3-642-12686-4_5.

[7] T. Hong, Short term electric load forecasting, Ph.D. thesis, North Carolina State University (2010).

[8] P. Gaillard, Y. Goude, R. Nedellec, Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting, International Journal of forecasting 32 (3) (2016) 1038–1050. doi:10.1016/j.ijforecast.2015.12.001.

[9] V. Dordonnat, A. Pichavant, A. Pierrot, GEFCom2014 probabilistic electric load forecasting using time series and semi-parametric regression models, International Journal of Forecasting 32 (3) (2016) 1005–1011. doi:10.1016/j.ijforecast.2015.11.010.

[10] N. Cesa-Bianchi, G. Lugosi, Prediction, learning, and games, Cambridge university press, 2006. doi:10.1017/CBO9780511546921.008.

[11] S. Haben, G. Giasemidis, A hybrid model of kernel density estimation and quantile regression for GEFCom2014 probabilistic load forecasting, International Journal of Forecasting 32 (3) (2016) 1017–1022. doi:10.1016/j.ijforecast.2015.11.004.

[12] E. Mangalova, O. Shesterneva, Sequence of nonparametric models for GEFCom2014 probabilistic electric load forecasting, International Journal of Forecasting 32 (3) (2016) 1023–1028. doi:10.1016/j.ijforecast.2015.11.001.

[13] F. Ziel, B. Liu, Lasso estimation for GEFCom2014 probabilistic electric load forecasting, International Journal of Forecasting 32 (3) (2016) 1029–1037. doi:10.1016/j.ijforecast.2016.01.001.

[14] D. Goldberg, Genetic algorithms in search, optimization and machine learning, Massachusetts: Addison-Wesley, 1989.

[15] J. H. Holland, Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, MIT press, 1992. doi:10.1086/418447.

[16] D. Broadhurst, R. Goodacre, A. Jones, J. J. Rowland, D. B. Kell, Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry, Analytica Chimica Acta 348 (1-3) (1997) 71–86. doi:10.1016/s0003-2670(97)00065-2.

[17] R. Leardi, A. L. Gonzalez, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemometrics and intelligent laboratory systems 41 (2) (1998) 195–207. doi:10.1016/s0169-7439(98)00051-3.

[18] R. M. Jarvis, R. Goodacre, Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data, Bioinformatics 21 (7) (2005) 860–868. doi:10.1093/bioinformatics/bti102.

[19] R. S. Mariano, Comparing Predictive Accuracy AU - Diebold, Francis X., Journal of Business & Economic Statistics 13 (3) (1995) 253–263. doi:10.1080/07350015.1995.10524599.

[20] S.-H. Lai, T. Hong, When one size no longer fits all: Electric load forecasting with a geographic hierarchy, SAS White Paper.

[21] R. Fletcher, Practical methods of optimization, John Wiley & Sons, 1987. doi:10.1002/9781118723203.

[22] D. C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, Mathematical programming 45 (1-3) (1989) 503–528. doi:10.1007/BF01589116.

[23] S. Li, M. Tan, Tuning SVM parameters by using a hybrid CLPSOBFGS algorithm, Neurocomputing 73 (10-12) (2010) 2089–2096. doi:10.1016/j.neucom.2010.02.013.

[24] W. Zheng, P. Bo, Y. Liu, W. Wang, Fast B-spline curve fitting by L-BFGS, Computer Aided Geometric Design 29 (7) (2012) 448–462. doi:10.1016/j.cagd.2012.03.004.