Online aggregation of conformal predictive systems

Vladimir G. Trunov TRUNOV@IITP.RU
Institute for Information Transmission Problems
Vladimir V. V'yugin VYUGIN@IITP.RU
Institute for Information Transmission Problems

Editor: Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

Abstract

The problem of online probabilistic forecasting is considered. Probabilistic forecasts are obtained as a result of the application of conformal predictive systems. The conformal predictive system is a novel method for obtaining reliable predictions which are based on point forecasts of the regression algorithm. The paper considers the case when at each moment of time several competing conformal predictive systems (experts) give their predictions in the form of probability distribution functions. Probabilistic forecasts of the experts are combined by an aggregation algorithm into one probabilistic forecast at each step of the forecasting process, while expert forecasts can be used partially.

The developed methods are used to solve the well-known problem of predicting the load of an electrical network online. Numerical experiments have shown the agreement of predictions with real data.

Keywords: Conformal prediction, Predictive distributions, Split conformal predictive systems, Aggregating of predictive distributions, Electrical load forecasting.

1. Introduction

Probabilistic predictions are important in many applications since the predictive distribution function describes the uncertainty of the prediction and also provides the ability to calculate the probabilities of any event associated with the predicted parameter.

We consider methods for predicting the test labels y of objects x, where $x \in \mathcal{R}^k$, (in the simplest case $k = 1$) and $y \in \mathcal{R}$. It is assumed that "object-label" pairs (x, y) are generated by some probability source (distribution), moreover, the pairs (x, y) are independent and identically distributed (iid). A weaker hypothesis on data exchangeability can also be used as the main assumption. We refer to such an assumption as to main assumption or main hypothesis Vovk et al. (2005). The specific form of this probability distribution may be unknown to us and will not be used in what follows.

There are a large number of methods for point, interval and probabilistic forecasting. The first part of this work is related to improving the quality and reliability of known methods, based on a recently proposed non-parametric approach called conformal predicting systems Vovk et al. (2019b). Conformal predictive systems are designed to make reliable probabilistic predictions for test labels basing only on the relationship between the current and past point forecasts of an arbitrary algorithm.

In papers Vovk et al. (2018a), Vovk et al. (2018b), Vovk et al. (2018b), Vovk et al. (2019b), Vovk et al. (2020b) split conformal predictive systems were introduced, as well

as cross-conformal predictive systems Vovk et al. (2020b). The first type system splits the training data into two parts - training and calibration samples, the second type system uses a cross-validation process to calculate the corresponding statistics. We will use systems of the first type.

Any conformal predictive system is constructed as follows. Training data is divided into training and calibration samples. The training sample is used to construct a prediction rule (algorithm) that, using the object x, presents a point prediction of its label y. The calibration sample allows, based on a comparison of a point forecast with past forecasts and known outcomes, to rebuild the algorithm forecast into (a predictive) probability distribution. This makes it possible to assess the inaccuracy and uncertainty for the point forecast. For example, using this probability distribution, one can build confidence interval predictions for y.

The training sample is used only at the initial stage of training — for constructing a regression algorithm. This algorithm will be used at the following steps to construct conformal predictive systems.

Conformal predictive system will be constructed online basing on point forecasts of the regression algorithm. At each time step $t = 1, 2, \ldots$ the system receives the testing object x_t, for which it is necessary to determine the distribution function of its label y_t. To do this, the calibration sample $\tilde{z}_1^m = (\tilde{x}_1, \tilde{y}_1), \ldots, z_m = (\tilde{x}_m, \tilde{y}_m)$ is used, consisting of the observed objects and their labels, the corresponding forecasts are also used by algorithm for these labels.

The probability distribution for the test label is constructed using a special statistic – a measure of conformity (or nonconformity) of the pair (x, y) with respect to the calibration sample $\tilde{z}_1^m$.

In order to achieve more accurate forecasts, the approach of Prediction with Expert Advice is used Vovk (1998), Vovk (2001).

At the training stage, the training part of the time series is divided into sections of homogeneity, which are determined based on the properties of the subject area. This data split used is essentially a Mondrian categories, see Vovk et al. (2005), Boström et al. (2021).

The Mondrian partition breaks the data into regions of homogeneity, within which the main hypothesis gets more evidence. Mondrian conformal predictive distributions based on Mondrian categories significantly outperforms the use of standard conformal predictive distributions, see Boström et al. (2021), Vovk (2022).

Mondrian categories will be extended to "fuzzy sets" when aggregating conformal distributions. Although each expert is trained in his area of expertise, we will partially use his predictions in neighboring areas as well, bearing in mind that his predictive ability declines gradually outside of his area of expertise. It is also necessary to take into account the fact that we use conformal predictive systems for different Mondrian partitions (seasonal partitions, additional time-of-day partitions).

Thus, at each time point, we can have probabilistic forecasts from several experts at once and have to aggregate them, taking into account the degree of their competence. Numerical experiments show that the proposed aggregation algorithm successfully copes with this task.

Main contribution of this paper is follows:

- Two methods for constructing online Mondrian conformal predictive systems are proposed and tested.

- A method for online aggregating of conformal predictive systems in the prediction of expert advice framework using experts' competence levels is presented.
- The algorithm has been developed for obtaining probabilistic forecasts online based on the proposed methods.
- Empirical support for the proposed approach is provided, showing that predictive performance of aggregating algorithm, as measured by CRPS loss function, may be improved compared to the individual experts.
- Based on real data on the load of the electrical network, a comparative analysis of the effectiveness of these and previously proposed forecasting methods was carried out. To illustrate the proposed methods for constructing predictive models and their aggregation, real data from the problem of predicting loads in an electrical network (see Devaine et al. (2013), Tao Hong (2016)) are used.

Section 2.1 contains the main concepts and definitions of the method of conformal predictions. Section 2.2 presents some details of the aggregating algorithm of Vovk (1998), as well as its extension to the case of probabilistic forecasts by V'yugin and Trunov (2019a). In Section 2.3 we generalize the aggregating algorithm for the case when expert predictions are provided with levels of competence. In this section the concept of discounted regret is introduced, its upper bound is obtained.

The effectiveness of the proposed methods is demonstrated by the results of numerical experiments given in section 4. The data of the Global Energy Forecasting - Competition 2014 (GEFCOM 2014, Track Load) used for the experimental study, a detailed description is given in Section 4 below.

2. Basic concepts and methods

2.1. Conformal Predictions

Let $\mathcal{X}$ be a measurable space, which we will call the object space. The observation space is defined as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$; its element $z = (x, y)$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, is interpreted as an observation consisting of the object x and its label y. The task of "supervised learning" is that, given training data consisting of observations $z_i = (x_i, y_i)$, $i = 1, \dots, n$ and a new (test) object $x_{n+1} \in \mathcal{X}$, predict the corresponding label y_{n+1}.

The classical formulation of the method of conformal predictions is given by Vovk et al. (2005), where conformal predictors use previous data to determine the size of confidence set Γ^{ϵ}_n, where $0 < \epsilon < 1$, for new predictions. Let there be an algorithm f that generates predictions $\hat{y}_{n+1}$. Then the described method makes it possible to obtain the set Γ^{ϵ}_n of predictions, which also contains the true value y_{n+1} with significance level $1 - \epsilon$. This method can be applied to any machine learning algorithm. The predictions of the algorithm have the property of validity if, with an increase in the number of outcomes (x_n, y_n) and the corresponding predictions $\hat{y}_{n+1} = f_{z_1^n}(x_{n+1})$, with the probability 1, the error rate will tend to ϵ.[1]

1. An error is a forecast result when $y_{n+1} \notin \Gamma^{\epsilon}_n$.

A conformity measure is a measurable function $A : \cup_{n=1}^{\infty} Z^{n+1} \to \mathcal{R}$, invariant under permutations of training observations: for any n and for any permutation π of the set $\{1, \dots, n\}$ for any sequence $z_1^n = z_1, \dots z_n \in Z^n$ and for any $z_{n+1} \in Z$,

$$A(z_1, \dots, z_n, z_{n+1}) = A(z_{\pi(1)}, \dots, z_{\pi(n)}, z_{n+1}).$$

The interpretation of the conformity measure A is that the value of $A(z_1, \dots, z_n, z_{n+1})$ measures how much a new observation z_{n+1} is "similar" to previously received observations $(z_1, \dots, z_n)$.

An example of the conformity measure for the case of regression (when $\mathcal{Y} = \mathcal{R}$) is

$$A(z_1, \dots, z_n, x_{n+1}) = y - \hat{y}_{n+1}, \tag{1}$$

where $\hat{y}_{n+1}$ is the label prediction calculated by the considered algorithm based on the known values of $z_1, \dots, z_n, x_{n+1}$ and y is the testing label.

Define conformity counters

$$\alpha_i^y = A(z_1, \dots, z_{i-1}, z_{i+1}, \dots, (x_{n+1}, y), z_i) \text{ for } i = 1, \dots, n, \tag{2}$$

$$\alpha_{n+1}^y = A(z_1, \dots, \dots, z_n, (x_{n+1}, y)). \tag{3}$$

The conformal transform corresponding to the conformity measure A is defined as

$$C^A(z_1^n, (x_{n+1}, y)) = \frac{1}{n+1} |\{i : 1 \le i \le n+1, \alpha_i^y \le \alpha_{n+1}^y\}|, \tag{4}$$

where $|D|$ denotes the number of elements of a finite set D.

It is easy to see that under the assumption that all pairs $z_i = (x_i, y_i)$ and x_{n+1} are generated independently by the some probability distribution P,

$$P\{C^A(z_1^n, (x_{n+1}, y)) \le \epsilon\} \le \epsilon. \tag{5}$$

Indeed, the left side of the inequality (5) is the probability that the number α_{n+1}^y is among the subset of the largest numbers of the set $\alpha_1^y, \dots, \alpha_{n+1}^y$, the fraction $\le \epsilon$. Since all orderings of this set are equally likely, the probability of the event (5) does not exceed ϵ.

The goal is to construct a probability distribution function based on (4), i.e., this value must be uniformly distributed. In this case, in (5) should be equality. To do this, the following modification of the definition (4) was carried out by Vovk et al. (2005). Let $\tau \in [0, 1]$ be a uniformly distributed random variable independent of all z_i.

A randomized conformal transform corresponding to the conformity measure A is defined as

$$C^A(z_1^n, (x_{n+1}, y), \tau) = \frac{1}{n+1} |\{i : 1 \le i \le n+1, \alpha_i^y < \alpha_{n+1}^y\}| + \frac{\tau}{n+1} |\{i : 1 \le i \le n+1, \alpha_i^y = \alpha_{n+1}^y\}|. \tag{6}$$

Vovk et al. (2005) has proved that the predictive system (6) is calibrated in probability:

$$P\{C^A(z_1^n, (x_{n+1}, y), \tau) \le \epsilon\} = \epsilon, \tag{7}$$

where P is the combined probability in z_i, x_{n+1} and $\tau \sim U$ (uniform distribution on the interval $[0,1]$).

The property (7) can be used to build prediction confidence sets.[2]. Any y such that

$$C^A(z_1^n, (x_{n+1}, y), \tau) > \epsilon, \tag{8}$$

is considered as a forecast value with $1-\epsilon$ significance level: $y \in \Gamma_n^\epsilon$. The condition (8) (and the definition (6)) means that the value of the conformity counter of the test pair (x, y) is not less than the values of conformity counters of the fraction ϵ of all pairs of the learning sample.

A distribution predicted for the test label must have the property of validity Vovk et al. (2019b), in other words, this means that the distribution must have statistical compatibility with implementations, i.e., reflect the true state of affairs. The main assumption to ensure such validity is that the elements of the calibration sample and the tested value x_t be independently and identically distributed (iid) with respect to some overall probability distribution, which we may not know.[3]

Vovk et al. (2005) has proved that the quantities $p_n = C^A(z_1^n, (x_{n+1}, y), \tau)$ are independently and identically distributed over P and U, which, together with (7), ensures the validity of predictions – according to the law of large numbers, with probability 1 the proportion of errors in choosing y tends to ϵ in the limit.

Split conformal predictive systems. Finding confidence sets for conformal predictions using the rule (8), even in the case of a finite set $\mathcal{Y}$, is a computationally difficult problem. In Vovk et al. (2018a), Vovk et al. (2018b), Vovk et al. (2020a) split conformal prediction systems that are much more computationally efficient [4] have been introduced. In this formulation, the entire training set is divided into a training set $z_1^n = z_1, \ldots, z_n$ and a calibration set $\tilde{z}_1^m = \tilde{z}_1, \ldots, \tilde{z}_m$. Based on the training sample, a basic algorithm is built that can make predictions $\hat{y} = f_{z_1^n}(x)$ for every x. Based on the calibration sample, we define the conformity measure $A(z_1^n, (x, y))$ and the conformity counters α_i of elements of the calibration sample and the counter α^y of an arbitrary test pair (x, y):

$$\alpha_i = A(z_1^n, \tilde{z}_i), \text{ for } i = 1, \ldots, m. \tag{9}$$

$$\alpha^y = A(z_1^n, (x, y)). \tag{10}$$

The corresponding randomized conformal transform is defined as

$$C^A_{z_1^n, \tilde{z}_1^m, x, \tau} = \frac{1}{m+1} |\{i : 1 \le i \le m, \alpha_i < \alpha^y\}| + \frac{\tau}{m+1} |\{i : 1 \le i \le m, \alpha_i = \alpha^y\}|, \tag{11}$$

where τ is a random variable uniformly distributed over the interval $[0,1]$.

The split conformal predictive system (11) is also calibrated in probability in the sense of (7), see Vovk et al. (2020a).

2. A confidence set is an interval if the function (6) is monotonic in y
3. A hypothesis on data exchangeability can also be used as the main assumption.
4. Although less accurate in applications, as noted in these papers.

The split conformal predictive system (11) allows to construct a conformal probability distribution of predictions of the corresponding labels instead of confidence sets of predictions. The implementation of the scheme for constructing a probabilistic split conformal prediction in online mode is given in section 3 below.

2.2. Prediction with Expert Advice

In the previous section, methods for constructing probability distribution function for any expert were given. The problem arises of aggregating these distribution functions into a single resulting probability distribution function.

In this section, we give the necessary definitions and auxiliary results of the theory of predictions with expert advice. A detailed presentation and proofs of the main statements are given in Vovk (1998) and Vovk (2001).

Online learning and forecasting. Let Ω be a set of outcomes, Γ be a set of predictions and $\lambda(f, y)$ be a real non-negative loss function, where $f \in \Gamma$ and $y \in \Omega$.[5] Also, let $E = \{1, \ldots, N\}$ be the set of experts.

In the theory of predictions with expert advice, the process of online learning and forecasting is considered as a game with complete information. At each step $t = 1, 2, \ldots$, each expert $i \in E$ makes a prediction $f_{i,t} \in \Gamma$, after that, the forecaster makes his prediction $f_t \in \Gamma$. After the predictions have been made, an outcome $y_t \in \Omega$ is announced and each expert i incurs the loss $\lambda(f_{i,t}, y_t)$ and the forecaster incurs loss $\lambda(f_t, y_t)$. This sequence of actions is presented below as Protocol 1.

Protocol 1

FOR $t = 1, \ldots, T$

1. Experts present forecasts $f_{i,t}$, where $1 \le i \le N$.
2. Forecaster presents his prediction f_t.
3. An outcome y_t is revealed and the losses $\lambda(f_{i,t}, y_t)$ of the experts and the loss $\lambda(f_t, y_t)$ of Forecaster are calculated.

ENDFOR

Let $H_T = \sum_{t=1}^{T} \lambda(f_t, y_t)$ be the total (accumulated) loss of Forecaster and $L_T^i = \sum_{t=1}^{T} \lambda(f_{i,t}, y_t)$ – total loss of the expert i suffered for the first T steps. The difference $R_T^i = H_T - L_T^i$ is called regret with respect to the expert i, $R_T = H_T - \min_i L_T^i$ – regret with respect to the least loss (best) expert. The purpose of Forecaster is to make predictions in such a way as to minimize the regret.

Aggregating algorithms AA and WA. To aggregate the expert forecasts, we will apply a scheme proposed in Vovk (1990), Vovk (1998), Vovk (2001), Cesa-Bianchi and Lugosi (2006).

Each of the AA and WA algorithms assigns weights to experts depending on their accumulated losses. At the first step the initial weights are defined as $w_{i,1} = \frac{1}{N}$ for all i. At subsequent steps, weights $w_{i,t}$ are updated according to the rule:

$$w_{i,t+1} = w_{i,t} e^{-\eta \lambda(f_{i,t}, y_t)} \text{ for } t = 1, 2, \ldots, \tag{12}$$

5. The nature of these sets will be explained later.

where $\eta > 0$ is a learning parameter. These weights are normalized:

$$w^*_{i,t} = \frac{w_{i,t}}{\sum_{j=1}^{N} w_{j,t}}. \tag{13}$$

The AA algorithm uses the so-called superprediction function to build an aggregated forecast. According to Vovk (1998), a loss function is said to be η-mixable if for any probability distribution $\mathbf{q} = (q_1, \ldots, q_N)$ on the set E of all experts [6] and for any of their predictions $\mathbf{f} = (f_1, \ldots, f_N)$ there exists a prediction f such that

$$\lambda(f, y) \le g(y) \text{ for all } y, \tag{14}$$

Where

$$g(y) = -\frac{1}{\eta} \ln \sum_{i=1}^{N} e^{-\eta \lambda(f_i, y)} q_i \tag{15}$$

is a superprediction function.

We fix the rule for computing prediction f and denote

$$f = \text{Subst}(\mathbf{f}, \mathbf{q}). \tag{16}$$

The function Subst is called substitution function.

For the AA algorithm, at each step t, the superprediction function is defined as

$$g_t(y) = -\frac{1}{\eta} \ln \sum_{i=1}^{N} e^{-\eta \lambda(f_{i,t}, y)} w^*_{i,t}, \tag{17}$$

at the same time, Forecaster's prediction is $f_t = \text{Subst}(\mathbf{f}_t, \mathbf{w}^*_t)$, where $\mathbf{f}_t = (f_{1,t}, \ldots, f_{N,t})$ be experts forecasts and $\mathbf{w}^*_t = (w^*_{1,t}, \ldots, w^*_{N,t})$ be the set of their normalized weights.

Regret bound for AA. Let the loss function $\lambda(f, y)$ be η-mixable, where $\eta > 0$, $\mathbf{w}^*_t = (w^*_{1,t}, \ldots, w^*_{N,t})$ be the normalized weights and $\mathbf{f}_t = (f_{1,t}, \ldots, f_{N,t})$ be the the experts' forecasts at step t. Let also, Forecaster calculate his forecast $f_t = \text{Subst}(\mathbf{f}_t, \mathbf{w}^*_t)$. By(14) $\lambda(f_t, y_t) \le g_t(y_t)$ for all t, where $g_t(y)$ is defined by (17). Let $H_T = \sum_{t=1}^{T} \lambda(f_t, y_t)$ be Forecaster's total loss and $L^i_T = \sum_{t=1}^{T} \lambda(f_{i,t}, y_t)$ be total loss of any expert i. By definition $g_t(y_t) = -\frac{1}{\eta} \ln \frac{W_{t+1}}{W_t}$, where $W_t = \sum_{i=1}^{N} w_{i,t}$ and $W_1 = 1$. According to the weight update rule (12), we get $w_{i,t+1} = \frac{1}{N} e^{-\eta L^i_t}$. Reducing the terms equal in absolute value and opposite in sign, we obtain a time-independent upper bound

$$H_T \le \sum_{t=1}^{T} g_t(y_t) = -\frac{1}{\eta} \ln W_{T+1} \le L^i_T + \frac{\ln N}{\eta} \tag{18}$$

for an arbitrary expert i. Thus, there is the forecaster strategy that guarantees the upper bound $R_T \le \frac{\ln N}{\eta}$ for regret for all T.

6. i.e., $\sum_{i=1}^{N} q_i = 1$ and $q_i \ge 0$ for all i

2.3. Aggregation of probabilistic forecasts

Loss function CRPS. Let the set of outcomes in Protocol 1 be the interval $\Omega = [a, b]$ of the real line, where $a < b$, and the set of predictions Γ be the set of all probability distribution functions on this interval: $F : [a, b] \to [0, 1]$. The continuous ranked probability score (CRPS loss function) is defined as

$$\mathrm{CRPS}(F, y) = \int_a^b (F(u) - H(u - y))^2 du, \tag{19}$$

where $y \in [a, b]$ is an outcome and $H(x)$ is the Heaviside function: $H(x) = 0$ for $x < 0$ and $H(x) = 1$ for $x \geq 0$.

Consider a probability forecasting game with expert advice. At each step t, each expert $i \in \{1, \ldots, N\}$ presents a forecast which is a probability distribution function $F_{i,t}(u)$, Forecaster presents his prediction $F_t(u)$. After that, an outcome $y_t \in [a, b]$ is revealed and the experts and Forecaster suffer losses $\mathrm{CRPS}(F_{i,t}, y_t)$ and $\mathrm{CRPS}(F_t, y_t)$.

V'yugin and Trunov (2019a) (and V'yugin and Trunov (2022)) proved that the loss function $\mathrm{CRPS}(F, y)$ is η-mixable for $0 < \eta \leq \frac{2}{b-a}$ and η-exponentially convex for $0 < \eta \leq \frac{1}{2(b-a)}$ and, therefore, the regret bounds from Section 2.2 are valid. For $\eta = \frac{2}{b-a}$ Forecaster's prediction $F_t(u)$ is computed from expert's forecasts $F_{i,t}(u)$, where $1 \leq i \leq N$, according to the rule

$$F_t(u) = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^N w_{i,t}^* e^{-2(F_{i,t}(u))^2}}{\sum_{i=1}^N w_{i,t}^* e^{-2(1-F_{i,t}(u))^2}}, \tag{20}$$

where $w_{i,t}^* = \frac{w_{i,t}}{\sum_{j=1}^N w_{j,t}}$ are normalized weights and $w_{i,t+1} = w_{i,t} e^{-\frac{2}{(b-a)}\mathrm{CRPS}(F_{i,t}, y_t)}$.

For algorithm WA prediction rule (20) should be replaced with

$$F_t(u) = \sum_{i=1}^N w_{i,t}^* F_{i,t}(u), \tag{21}$$

where $w_{i,t}^* = \frac{w_{i,t}}{\sum_{j=1}^N w_{j,t}}$ are the normalized weights and $w_{i,t+1} = w_{i,t} e^{-\frac{1}{2(b-a)}\mathrm{CRPS}(F_{i,t}, y_t)}$. For algorithm AA regret bound is $\frac{b-a}{2} \ln N$, see V'yugin and Trunov (2019a). For WA regret bound is $2(b-a) \ln N$.[7]

2.4. AA for Experts with Competence Levels

In this section, we will somewhat expand the formulation of the prediction problem. Let at each time t forecasts $\mathbf{f}_t = (f_{1,t}, \ldots f_{N,t})$ experts are provided with levels of competence $\mathbf{p}_t = (p_{1,t}, \ldots, p_{N,t})$.

Competence level $p_{1,t}$ is a real number from the interval $[0, 1]$. If $p_{i,t} = 0$, then the corresponding expert "sleeps" at step t, i.e., his prediction is not taken into account by the aggregating algorithm. If $p_{i,t} < 1$, then this forecast $f_{i,t}$ will be used only partially with a discount, see V'yugin and Trunov (2019a).

Associate each level of competence $p_{i,t}$ with a probability distribution $\mathbf{p}_{i,t} = (p_{i,t}, 1 - p_{i,t})$ on a two-element set and. define an auxiliary randomized forecast of the "virtual" expert i:

$$\tilde{f}_{i,t} = \begin{cases} f_{i,t} \text{ with probability } p_{i,t}, \\ f_t \text{ with probability } 1 - p_{i,t}, \end{cases}$$

7. Note that the regret bound for WA is four times worse, than the bound for AA. Results of experiments in Section 4 show that, the accumulative losses of the AA algorithm are less than those of the WA algorithm.

where f_t is Forecaster's prediction, which will be calculated later.

Our goal is to define the forecast f_t so that

$$e^{-\eta\lambda(f_t,y)} \geq \sum_{i=1}^{N} E_{\mathbf{p}_{i,t}}[e^{-\eta\lambda(\tilde{f}_{i,t},y)}]w_{i,t} \tag{22}$$

for every y. Here $E_{\mathbf{p}_{i,t}}$ is the expectation with respect to the probability distribution $\mathbf{p}_{i,t}$. Also $w_{i,t}$ is the weight of expert i at step t.

Let's write the inequality (22) in a more detailed form:

$$e^{-\eta\lambda(f_t,y)} \geq \sum_{i=1}^{N} E_{\mathbf{p}_{i,t}}[e^{-\eta\lambda(\tilde{f}_{i,t},y)}]w_{i,t} = \tag{23}$$

$$\sum_{i=1}^{N} p_{i,t}w_{i,t}e^{-\eta\lambda(f_{i,t},y)} + e^{-\eta\lambda(f_t,y)}\left(1 - \sum_{i=1}^{N} p_{i,t}w_{i,t}\right) \tag{24}$$

for all ω. Thus, the inequality (22) is equivalent to the inequality

$$e^{-\eta\lambda(f_i,y)} \geq \sum_{i=1}^{N} w^*_{i,t}e^{-\eta\lambda(f_{i,t},y)}, \tag{25}$$

where

$$w^*_{i,t} = \frac{p_{i,t}w_{i,t}}{\sum_{j=1}^{N} p_{j,t}w_{j,t}}. \tag{26}$$

We use the rule (16) to calculate a forecast of the AA algorithm: $f_t = \text{Subst}(\bar{f}_t, \mathbf{w}^*_t)$. Then (25) is equivalent to (23). Here Subst is a substitution function, $\mathbf{w}^*_t = (w^*_{i,1}, \ldots, w^*_{i,N})$ and $\bar{f}_t = (f_{1,t}, \ldots f_{i,N})$.

Let $h_t = \lambda(f_t, y_t)$ be the Forecaster's loss at step t, Virtual Expert i suffers the loss $\hat{l}_{i,t} = E_{\mathbf{p}_{i,t}}[\lambda(\tilde{f}_{i,t}, y_t)]$.

Protocol 1 for AA with competence levels is modified as follows, see V'yugin and Trunov (2019a):

Protocol 2

FOR $t = 1, \ldots, T$

1. Get $f_{i,t}$ expert predictions and competence levels $p_{i,t}$, where $1 \leq i \leq N$.
2. Define Forecaster's prediction $f_t = \text{Subst}(\mathbf{f}_t, \mathbf{w}^*_t))$, where $\mathbf{w}^*_t = (w^*_{1,t}, \ldots, w^*_{N,t})$ are normalized weights defined by
$$w^*_{i,t} = \frac{p_{i,t}w_{i,t}}{\sum_{j=1}^{N} p_{j,t}w_{j,t}}.$$
3. Get the true value of the outcome y_t and calculate the loss $l_{i,t} = \lambda(f_{i,t}, y_t)$ of experts and the loss of Forecaster $\lambda(f_t, y_t)$.
4. Update the experts' weights:
$$w_{i,t+1} = w_{i,t}e^{-\eta(p_{i,t}\lambda(f_{i,t},y_t)+(1-p_{i,t})\lambda(f_t,y_t))}. \tag{27}$$

ENDFOR

Since by definition the loss of a virtual expert is $\hat{l}_{i,t} = p_{i,t}l_{i,t} + (1 - p_{i,t})h_t$, will be $h_t - \hat{l}_{i,t} = p_{i,t}(h_t - l_{i,t})$. Let's call this value the discounted regret with respect to the expert i at the step t. We will measure performance of our algorithm using the total discounted regret with respect to the expert i.

Theorem 1 *For any $1 \le i \le N$ the upper bound of the total discounted regret relative to expert i:*

$$\sum_{t=1}^{T} p_{i,t}(h_t - l_{i,t}) \le \frac{\ln N}{\eta}. \tag{28}$$

Proof. From the convexity of the exponent and the inequality (22) we get

$$e^{-\eta\lambda(f_t,y)} \ge \sum_{i=1}^{N} e^{-\eta E_{\mathbf{p}_{i,t})}[\lambda(\tilde{f}_{i,t},y)]} w^*_{i,t} = \sum_{i=1}^{N} e^{-\eta\hat{l}_{i,t}} w^*_{i,t}. \tag{29}$$

Let $m_t = -\frac{1}{\eta}\ln\sum_{i=1}^{N} w^*_{i,t}e^{-\eta\hat{l}_{i,t}}$ .By (29) $h_t \le m_t$. Let's rewrite the rule (27) as

$$w_{i,t+1} = w_{i,t}e^{-\eta\hat{l}_{i,t}}. \tag{30}$$

Recall that $W_T = \sum_{t=1} w_{i,t}$, $W_1 = 1$ and $m_t = \frac{1}{\eta}\ln\frac{W_{t+1}}{W_t}$. Just like in (18), using (29) and (30), we get

$$\sum_{t=1}^{T} h_t \le \sum_{t=1}^{T} m_t = -\frac{1}{\eta}\ln W_{T+1} \le \sum_{t=1}^{T} \hat{l}_{i,t} + \frac{\ln N}{\eta}$$

for any i. Since $h_t - \hat{l}_{i,t} = p_{i,t}(h_t - l_{i,t})$, we get the inequality (28). △

3. Online construction of conformal distributions

To build and calibrate expert strategies, the entire array of historical data, consisting of pairs (x, y), where x is the temperature, y is the network load, is divided by pairwise non-overlapping intervals of time segments (season, time of day), which will also be called the areas of competence of the respective experts.

The data split used is essentially a Mondrian categories. The Mondrian partition breaks the data into regions of homogeneity, within which the main hypothesis gets more evidence.[8]

Experts training. The entire array of historical data, consisting of pairs (x_t, y_t), where x_t is the temperature, y_t is the network load at time t, is divided by pairwise non-overlapping intervals of time segments (season, time of day), that are the areas of competence of the respective experts. In each area of competence, the data is divided into a training sample $z_1^n = z_1, \ldots, z_n$, and a calibration sample $\tilde{z}_1^m = \tilde{z}_1, \ldots, \tilde{z}_m$, where $z_i = (x_{t_i}, y_{t_i})$ for $1 \le i \le n$ and $\tilde{z}_i = (\tilde{x}_{t_i}, \tilde{y}_{t_i})$ for $1 \le i \le m$.

Each expert is trained on its own training set z_1^n [9] and will be in what follows calibrated on the elements of the $\tilde{z}_1^m$ set or on its extension, which will be determined online. [10]

8. The conformal predictive distributions based on Mondrian categories significantly outperforms the use of standard conformal predictive distributions, see Boström et al. (2021), Vovk (2022).
9. Based on this set, a regression equation is constructed that allows given x to calculate a point prediction y.
10. For this sample, based on the point predictions of the regression algorithm, as a result of online calibration, a conformal predictive system will be constructed that, given x, produces a probability distribution function of y.

Each expert will be trained and calibrated on elements of its area of competence (Mondrian category). Three types of splits are used: all data, splits by season (winter, spring, summer, autumn), split by time of day within a season (morning in winter, morning in spring, afternoon in winter, day in spring, etc.). AnyTime expert is trained and calibrated from all data, four seasonal experts are determined from seasonal data, the remaining 16 experts correspond to the times of the day.

Experts online calibration. The conformal calibration procedure defines a split conformal forecasting system. At each time step t, we get the testing object x_t, for which a probability distribution function of the consumption value y_t should be constructed using a calibration sample $\tilde{z}_1^m = (\tilde{x}_1, \tilde{y}_1), \ldots, \tilde{z}_m = (\tilde{x}_m, \tilde{y}_m)$, whose elements belong to the expert's area of competence (Mondrian category).

Two methods for forming calibration sample will be used. With the CP method, a part of the training sample is allocated, which serves as a calibration sample at all subsequent steps.[11]

With the CP+ method, the initial calibration sample is replenished at each step with new pairs from the area of competence observed at time t by the expert.[12]

Online aggregation of the experts' conformal distribution functions. In this paper, when aggregating the predictive distributions of experts, we somewhat expand the concept of Mondrian partitioning – by specifying a partition using real values p_i, we define fuzzy sets in which conformal predictive systems are applied and aggregated. The conducted comparative experiments (below) show that in this way we achieve more accurate results in forecasting.

In Section 4 (below) we will present the results of numerical experiments with real data. Below is a modification of Protocol 2, Algorithm 3, that will be used in these experiments.

Algorithm 3

FOR $i = 1, \ldots N$ **Preprocessing loop*
Using the ith training sample z_1^n [13], we build a regression rule (algorithm) $y = f_i(x)$.
ENDFOR **End of preprocessing loop*

Define $w_{i,1} = \frac{1}{N}$ for $1 \le i \le N$.
FOR $t = 1, \ldots, T$ **Main Loop*

1. We get the testing object x_t and define the probabilistic forecasts of experts – probability distribution functions $F_{i,t}(y)$ for $i = 1, \ldots N$.

   FOR $i = 1, \ldots N$ **Construction of the experts' conformal distributions.*[14]

   Let us fix the calibration sample $\tilde{z}_1^m = \tilde{z}_1, \ldots, \tilde{z}_m$ from the area of competence of the corresponding expert i, $\tilde{z}_s = (\tilde{x}_s, \tilde{y}_s)$ for $1 \le s \le m$.

   We use the conformity measure

   $$A(z_1^n, (x, y)) = y - \hat{y},$$

   where $\hat{y} = f_i(x)$ is the label prediction computed by the regression algorithm.

   Calculate the conformity counters α_s for $s = 1, \ldots, m$: $\alpha_s = A(z_1^n, (\tilde{x}_s, \tilde{y}_s))$ and arrange them in ascending order:

   $$\alpha_{(1)} < \cdots < \alpha_{(k)}.$$

   Let $n_j = |\{s : \alpha_s = \alpha_{(j)}\}|$ for $j = 1, \ldots, k$.

11. At the same time, the main assumption is preserved that the elements of the calibration sample and the test value must be independently and identically distributed distributed with respect to some probability distribution on pairs (x, y) corresponding to the the expert's area of competence. The specific form of this distribution is not taken into account.
12. This method makes it possible to take into account possible local violations of the basic assumption.
13. more precisely, its training part, from the area of competence of the expert i
14. Here we follow ideas from Vovk et al. (2020a).

Define also $m_j = \sup\{y : \alpha^y < \alpha_{(j)}\}$ and $M_j = \inf\{y : \alpha^y > \alpha_{(j)}\}$, where $\alpha^y = A(z_1^n, (x, y))$. Define the predictive conformal probability distribution function:

$$Q_{z_1^n, \tilde{z}_1^m, x_t, \tau}(y) = \begin{cases} \frac{\tau}{m+1} \text{ if } y < m_1, \\ \frac{n_1 + \dots + n_{j-1} + \tau n_j + \tau}{m+1} \text{ if } m_j < y < M_j, j = 1, \dots k, \\ \frac{n_1 + \dots + n_j + \tau}{m+1} \text{ if } M_j < y < m_{j+1}, j = 1, \dots k-1, \\ \frac{n_1 + \dots + n_k + \tau}{m+1} = \frac{m+\tau}{m+1} \text{ if } y > M_k. \end{cases}$$

It was proved by Vovk et al. (2020a) that, under some mild assumptions, the function Q is a probability distribution function.

Denote $F_{i,t}(y) = Q_{z_1^n, \tilde{z}_1^m, x_t, \tau}(y)$ the conformal probability distribution function of the expert i.

ENDFOR **End of loop for constructing the experts' conformal distributions*

2. *Aggregation of the experts' probability distribution functions.*

   We get the competence levels $p_{i,t}$, where $1 \le i \le N$. Define the probability distribution function of Forecaster by the rule

$$F_t(y) = \frac{1}{2} - \frac{1}{4} \ln \frac{\sum_{i=1}^N w_{i,t}^p e^{-2(F_{i,t}(y))^2}}{\sum_{i=1}^N w_{i,t}^p e^{-2(1-F_{i,t}(y))^2}} \tag{31}$$

   for AA algorithm, or by the rule

$$F_t(y) = \sum_{i=1}^N w_{i,t}^p F_{i,t}(y) \tag{32}$$

   for WA algorithm, where

$$w_{i,t}^p = \frac{p_{i,t} w_{i,t}}{\sum_{j=1}^N p_{j,t} w_{j,t}}.$$

3. Observe outcome y_t and compute losses $\mathrm{CRPS}(F_{i,t}, y_t)$ of the experts $1 \le i \le N$, as well as the loss $\mathrm{CRPS}(F_t, y_t)$ of Forecaster.

4. Update weights of the experts $1 \le i \le N$

$$w_{i,t+1} = w_{i,t} e^{-\eta(p_{i,t}\mathrm{CRPS}(F_{i,t}, y_t) + (1-p_{i,t})\mathrm{CRPS}(F_t, y_t))}, \tag{33}$$

   where $\eta = \frac{2}{b-a}$ for AA and $\eta = \frac{1}{2(b-a)}$ for WA.

ENDFOR **End of the main loop*

4. Probabilistic forecasting of hourly electrical loads

The data of the Global Energy Forecasting - Competition 2014 (GEFCOM 2014, Track Load) which was served as the material for the experimental study was held on the Kaggle platform (Tao Hong (2016)).

The goal of GEFCom2014-L was to evaluate quantiles (more precisely, all 100 percentiles) for the probability distribution of hourly electrical loads. At the same time, the horizon forecast varied over a wide range from one hour to one month. The main block of the training sample includes data on hourly electrical loads for the period from January 2005 to the end of 2010 and data on hourly temperature measurements at 25 meteorological stations from January 2001 to September 2010 (for

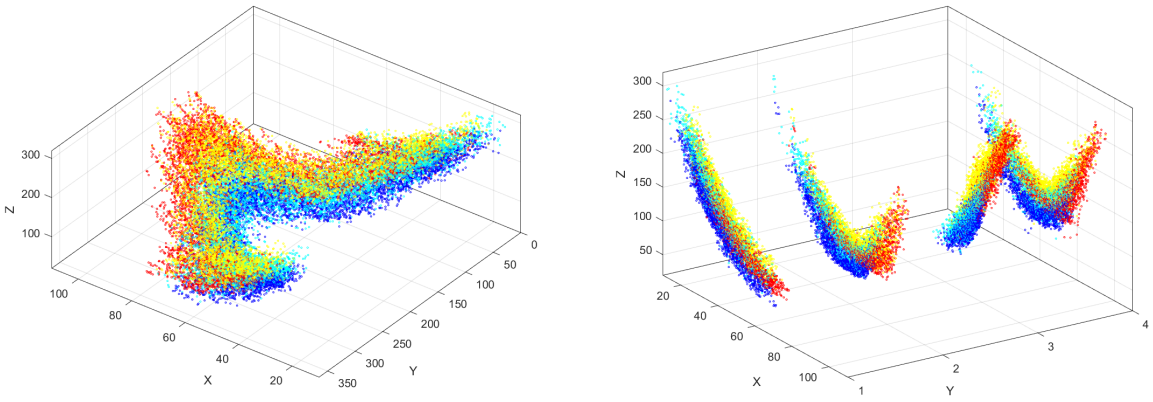

Figure 1: 3D "Temperature–Electrical Load" scatterplots showing the relationship between hourly air temperature measurements and electrical load on the network. The left figure shows a scattering cloud points "Temperature", "Load" , "Day of the year" . The right figure shows the same data, but the scatterplot points from the left figure refer to the same season of the year (winter, spring, summer, autumn) "collapsed" into flat scatterplots, which differ in the shape of the clouds. The color indicates the periods of the day (night, morning, afternoon, evening). The x-axis shows the temperature values, and the y-axis shows all the days of the year from 1 to 365.

117 months). The test sample includes data from January 1, 2011 to December 2011. The databases are available at $http://www.kaggle.com/datasets$.

In this work, we restrict ourselves to the data of the current temperature averaged over weather stations and its prehistory and calendar indicators (season of the year and time of day). In Fig. 1 3D "Temperature–Electrical Load" scatterplots showing the relationship between hourly air temperature measurements and electrical load on the network are presented.

4.1. Numerical experiments

In this section, we present the methodology that used for probabilistic forecasting of the electrical load on the electrical network.

Experts training. Taking into account the type of 3D-scatterplots (see Fig. 1), the training set was divided into subsamples corresponding to the seasons (winter, spring, summer, autumn) and time of day (morning, afternoon, evening, night).

The area of competence of any expert i at time t is determined by levels of competence $p_{i,t}$. At the training stage of the expert i, only those points t are included in the training sample, where $p_{i,t} = 1$. When forecasting, the scope of the expert extends to all points in time where his level of competence is greater than 0. Thus, each the expert competes with other experts working on overlapping intervals. In this case, the corresponding AA (or WA) algorithm for aggregating experts is used, taking into account their levels of competence.

Taking into account the analysis of diagrams scattering, shown in Fig. 1, the full training set was divided into subsamples, corresponding to the four seasons (winter, spring, summer, autumn) and the four time intervals (morning, afternoon, evening, night). Each of the subsamples was associated with a specialized expert, who formed his own probabilistic predictive model based on these data:

1.Reference model GMM – approximation of a two-dimensional point cloud using several Gaussian components, as was done in V'yugin and Trunov (2019a).

2.Methods CP (CP+) of conformal predictions of the experts which are based on point predictions of polynomial regression and subsequent use of calibration sample (constant - CP, or replenished - CP+) to build predictive probability distribution.

On the test sample, the use of specialized experts was regulated by their competence levels $p_{i,t}$. Since, when moving to the next calendar subsample, this expert gradually loses its predictive ability, the level of competence $p_{i,t}$ in its forecasts outside its area of competence decreased linearly from 1 to 0 in the extended part of the interval (see Fig. 2). Therefore, in the vicinity of the boundaries of the corresponding subsamples, there appear competing experts with different non-zero levels of competence. The aggregating algorithm (Algorithm 3) combines forecasts of the experts (probability distributions) taking into account these levels of competence.[15]

As can be seen from the scatterplots in Fig. 1, the season and time of day are manifested in the form of noticeable regular changes in the general appearance of scattering clouds and their relative position. The results of the conformal predictive systems (CP and CP+) are compared with the results of reference method constructing probability distributions from empirical data – the Gaussian mixture method (GMM).

Gaussian mixture method GMM. Gaussian mixture method at the training stage builds a probability distribution for each subsample based on a scatterplot steam "temperature – load". Probability distribution functions produced by the GMM method are aggregated using the AA and WA algorithms (Algorithm 3) in the probability distribution function of the resulting forecast.

Conformal prediction systems. The same training samples were used for construction and estimation of methods of conformal distributions CP (CP+). For each expert at the initial stage of training the expert's area of competence is divided into two parts – training sample and calibration sample. First on the training set a polynomial regression approximation is built, where cubic polynomials were used. Examples of point prediction of electrical loads can be found in V'yugin and Trunov (2019b). The calibration sample is used in the subsequent steps to build a predictive probability distribution.

Based on the calibration sample and on point forecasts, a system of conformal prediction is constructed. The calibration sample can remain constant in the process of forecasting or replenish at each step with new pairs from the expert's area of competence.

The following actions are performed online (see Algorithm 3):

At each step of the forecasting period based on current temperature forecasts and a calibration sample defined at the initial step, the probability distribution of the load value is constructed. This method is referred to as CP. Thus, with this method, the calibration sample is constant.[16]

In the CP+ method, the initial calibration sample is replenished at each step with new pairs from the observed part of the expert's area of competence, after which the probability distribution of the load value is constructed.[17]

Aggregation of expert probability distribution functions. The distribution functions produced by conformal predictive systems for each expert are aggregated using the AA and WA algorithms (Algorithm 3) into the distribution function of the resulting prediction.

15. As shown in V'yugin and Trunov (2019a), the use of extended areas of competence leads to more accurate predictions and a corresponding reduction in the cumulative loss of aggregation algorithms.
16. Here we use the assumption that the probability distribution generating data observable in any subset of an expert's area of expertise is the same.
17. Thus, small changes in the parameters of the generating distribution are taken into account.

4.2. Results of numerical experiments

To illustrate the proposed methods a particular problem was chosen, namely, predicting the probability distribution function of electrical loads one hour ahead based on the temperature forecast and current calendar parameters.[18]

An example of setting expert competence levels is shown in Fig. 2.

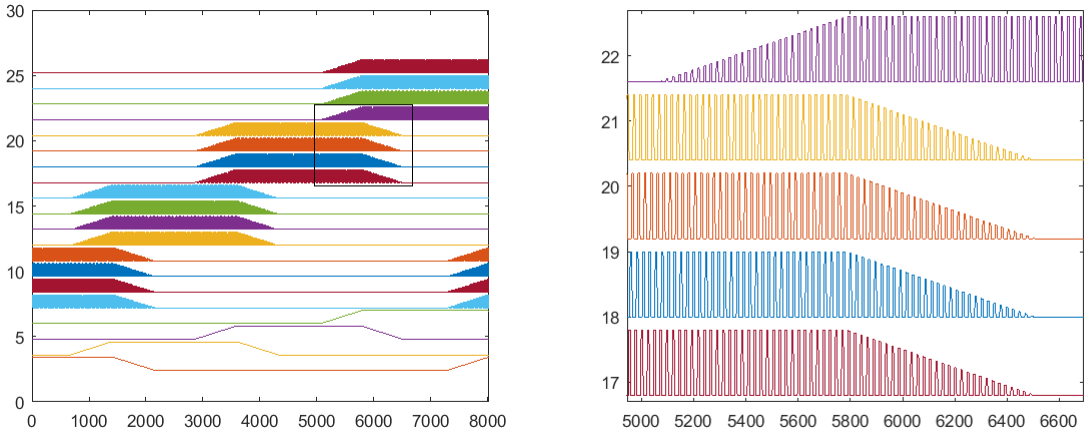

Figure 2: Left side of the figure: Competence levels for Experts 2-5 (seasonal experts) and 6-21 (Season, Time of day). Right side: Enlarged fragment of the drawing from the left side. Time steps are plotted along the x-axis, the competency value is plotted along the y-axis

In this experiment, the experts, which were built according to their areas of competence, make predictions at all points of the test sample, and the AA and WA algorithms aggregated these predictions in one case without taking into account their competence levels (see Fig. 3), and in the second case, taking into account the levels of competence (see fig. 4).

Fig. 3 shows graphs of average (over time) of accumulated CRPS losses: $t \to \frac{1}{t}\sum_{s=1}^{t} \mathrm{CRPS}(F_{i,s}, y_s)$, for experts 1–21, constructed by the CP method and the losses of two algorithms WA and AA, which are used without taking into account the expert's competence levels (i.e., $p_{i,t} = 1$ for all i and t.).

The figure 4 shows the average (over time) values of the accumulated CRPS losses of aggregators AA and WA for experts constructed by the GMM, CP and CP+ methods, applied taking into account the levels of expert competence.

The AA algorithm, when aggregating experts built by the CP and CP+ methods, leads to lower losses than the AA algorithm applied to experts built by the GMM method, and the use of the CP+ method for building experts leads to lower losses than the use of the CP method.

As before, for all methods of building experts throughout the test period, the loss of the aggregator AA is slightly less than the loss of the aggregator WA.

Discounted Regret curves of the AA Algorithm for all experts build by the CP Method are shown in Fig. 5.

The 3D image of the distribution density, built by the AA algorithm with the aggregation of experts built by the CP method, is shown in Fig.6.

18. Note that the above technology and the corresponding algorithms make it possible to calculate load forecasts at any future point in time, for which there is a temperature forecast.

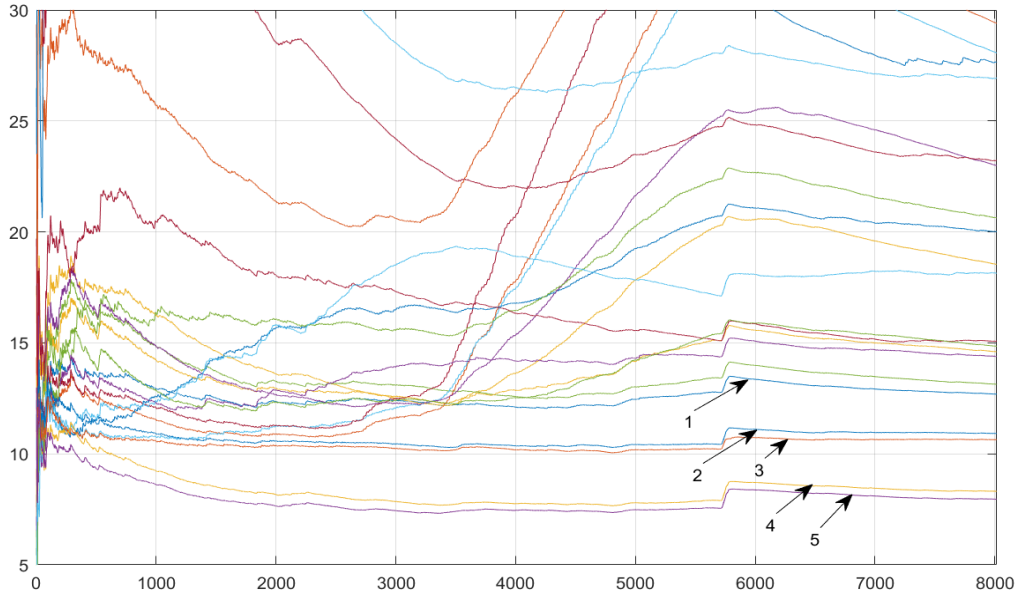

Figure 3: Average values of accumulated CRPS losses of experts, constructed by the CP method, and the loss of aggregators WA and AA of these experts, used without taking into account their levels of competence (marked with numbers 2 and 3). For comparison, the same figure shows the losses of aggregators WA and AA, built taking into account the levels of expert's competence (numbers 4 and 5). Number 1 marks the loss curve of the AnyTime expert. Time steps are plotted along the x-axis, the loss value is plotted along the y-axis

5. Conclusion

This paper presents the technology for probabilistic forecasting of loads in an electrical network, which uses methods for constructing conformal probabilistic forecasts (CP and CP+) and methods for aggregating these forecasts online. Experimental calculations are carried out on real data.

The results of forecasting by the CP and CP+ methods were compared with the results of the method for constructing probabilistic forecasts based on Gaussian GMM mixtures. To assess the quality of probabilistic forecasts, the continuous ranked probability scoring rule CRPS was used.

The paper presents the algorithm for predicting electrical network loads online using CP (CP+) conformal prediction systems. A comparative analysis of the effectiveness of the methods of conformal predictions CP and CP+ and the previously used method of Gaussian mixtures GMM for constructing the probability distribution functions of expert strategies has been carried out.

The GMM forecasting method for constructing experts uses an approximation of a two-dimensional point cloud "temperature – load" with the help of several Gaussian components. Conformal prediction systems (CP and CP+) do not directly use any approximation of data from the training sample, they build a predictive probability distribution only based on the relationship between the current and past point predictions of the regression algorithm.

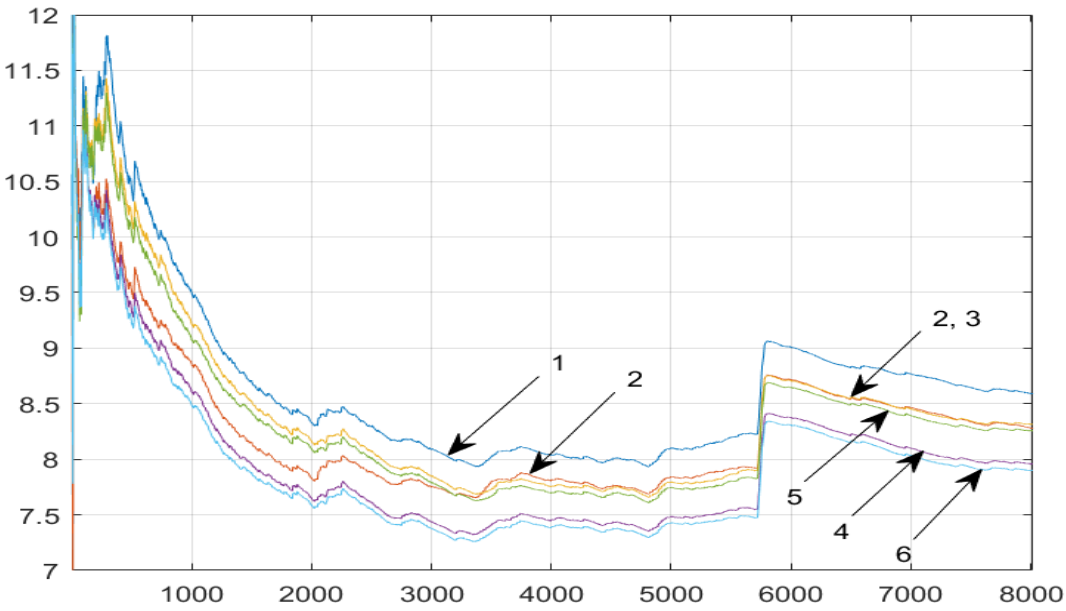

Figure 4: Average values of accumulated CRPS losses of aggregators AA and WA for experts constructed by GMM, CP and CP+ methods. Numbers mark the loss curves of 1-WA((GMM),2-AA(GMM), 3-WA(CP), 4-AA(CP), 5-WA(CP+), 6-AA(CP+) aggregators used with experts' levels of competence. Time steps are plotted along the x-axis, the loss value is plotted along the y-axis.

The numerical results show that the use of the CP and CP+ methods for the construction the expert forecasts leads to a smaller loss of averaged aggregated forecasts than method GMM and the use of the CP+ method for building experts leads to lower losses than the use of the CP method.

It is also shown that the AA aggregator has lower average losses than the WA aggregator over the entire test period, which is consistent with the corresponding regret bounds.

The conducted comparative experiments showed the highest accuracy of forecasts when using fuzzy sets of competence for aggregating experts. Perhaps even greater accuracy can be achieved by using fuzzy competence areas when constructing conformal predictor distributions, i.e., when calibrating predictions.

6. Acknowledgements

The authors would like to thank the anonymous reviewers for helpful comments.

References

H. Boström, U. Johansson and T. Löfström, 2021. Mondrian conformal predictive distributions. In Conformal and Probabilistic Prediction and Applications. PMLR, 152, pp. 24–38.

N. Cesa-Bianchi, G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

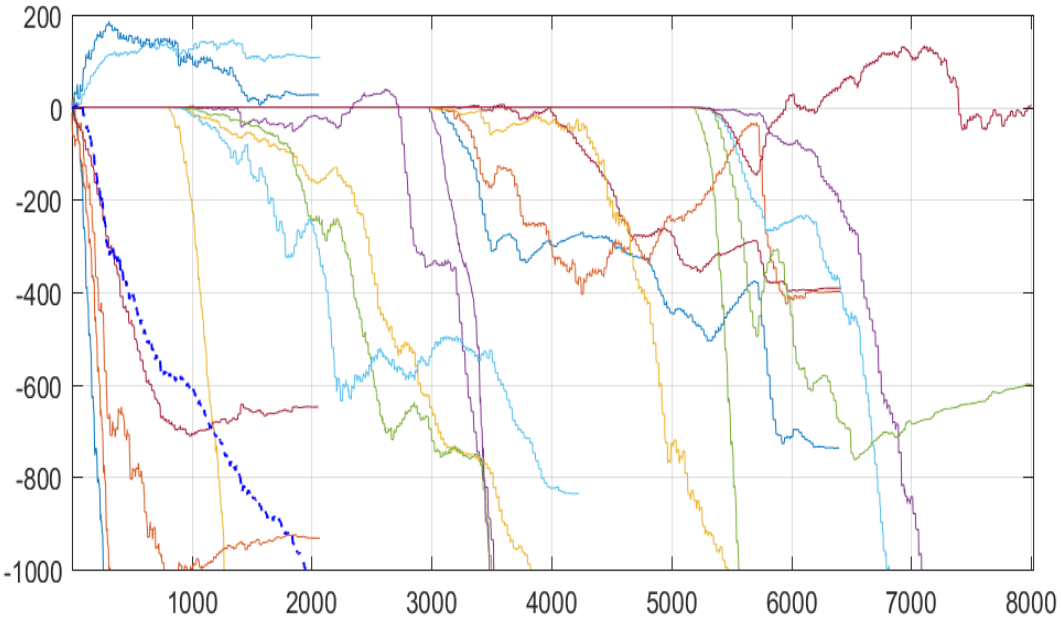

Figure 5: Discounted regrets of the AA algorithm with respect to Experts constructed by the CP method.

M. Devaine, P. Gaillard, Y. Goude, G. Stoltz. Forecasting electricity consumption by aggregating specialized experts. *Machine Learning*. 90(2): 231–260, 2013.

Tao Hong,, Pierre Pinson, Shu Fanc, Hamidreza Zareipour, Alberto Troccoli, Rob J. Hyndman. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting* 32: 896–913, 2016.

V. Vovk, Aggregating strategies. In M. Fulk and J. Case, editors, *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, 371–383. San Mateo, CA, Morgan Kaufmann, 1990.

V. Vovk, A game of prediction with expert advice. *Journal of Computer and System Sciences*. 56(2), 153–173, 1998.

V. Vovk. Competitive on-line statistics. *International Statistical Review* 69, 213–248, 2001.

V. Vovk, A. Gammerman, G. Shafer, 2005. Algorithmic learning in a random world. Springer Science and Business Media.

V. Vovk, I. Nouretdinov, V. Manokhin, A. Gammerman, Conformal predictive distributions with kernels, in: Braverman Readings in Machine Learning. Key Ideas from Inception to Current State. Springer. 2018, 103–121.

V. Vovk, I. Nouretdinov, V. Manokhin, A. Gammerman. Cross conformal predictive distributions. *Proceedings of Machine Learning Research, COPA* 91:37–51, 2018.

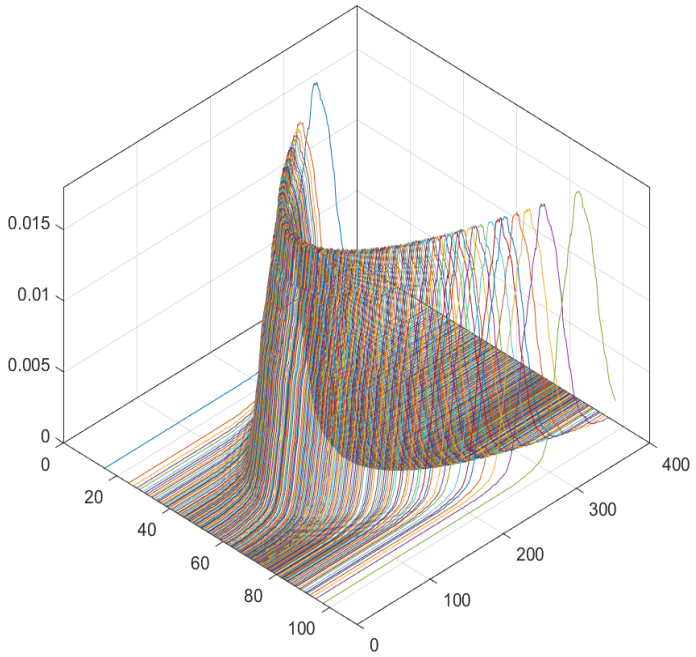

Figure 6: Densities of distributions built by the AA algorithm when aggregating Experts defined by the CP method.

V. Vovk, I. Nouretdinov, V. Manokhin, A. Gammerman. Computationally efficient versions of conformal predictive distributions. *Neurocomputing*. 397:292–308 (2020)

V. Vovk, J. Shen, V. Manokhin, Min-ge Xie. Nonparametric predictive distributions based on conformal prediction. *Machine Learning* 108(3), 445-474, 2019. https://doi.org/10.1007/s10994-018-5755-8

V. Vovk, J. Shen, V. Manokhin, Min-ge Xie. Nonparametric predictive distributions based on conformal prediction. *Machine Learning* 108(3), 445-474, 2019. https://doi.org/10.1007/s10994-018-5755-8

V. Vovk I. Petej, P. Toccaceli, A. Gammerman, E. Ahlberg, L. Carlsson, S. Line Conformal calibration. *Proceedings of Machine Learning Research* 128:1–16, 2020.

V. Vovk, I. Petej, P. Toccaceli, A. Gammerman, E. Ahlberg, L. Carlsson, Conformal calibration, in: Conformal and Probabilistic Prediction and Applications, *PMLR*. 84–99.

V. Vovk, 2022. Universal predictive systems. Pattern Recognition. 126: pp. 108536

V. V'yugin, V. Trunov. Online Learning with Continuous Ranked Probability Score, *Proceedings of Machine Learning Research* 105: 163–177, 2019.

V. V'yugin, V. Trunov. Online aggregation of unbounded losses using shifting experts with confidence. *Machine Learning*, 108(3): 425–444, 2019.

V. V'yugin, V. Trunov. Online aggregation of probability forecasts with confidence. Pattern Recognition Volume 121, January 2022, 108193.