# On applying the genetic algorithm in the aggregation of regression models

Jacek Stelmach[1]

**Abstract**

Aggregation of base models is one of the issues which determine the quality of forecasts regression models. Known literature describes mainly aggregation of classification models, usually recommending for regression models the arithmetic average of base models forecasts. The presented study adapts the weighted average method, often used in the aggregation of classification models. In the experiment, there were compared: aggregate functions with the weights calculated on the basis of selected measures of the quality of the base models - with weights determined by a genetic algorithm. The study was carried out using computer simulation, with both: simulated data and selected empirical data sets.

***Keywords:*** *aggregation model, genetic algorithm, weighted average*

***JEL Classification:*** C530

## 1. Introduction

Good features of aggregated models in regression analysis coming from the use of base models that are constructed in a variety of ways. One of the crucial issues that influences the quality of prediction capabilities is the way how the forecasts of $D_1, \ldots, D_M$ base models are aggregated into aggregated $D^*$ model. Known literature presents a number of aggregation functions that defines the aggregation rule of $M$ base models:

$$\hat{D}^*(\mathbf{x}) = \Psi(\hat{D}_1(\mathbf{x}), \ldots, \hat{D}_M(\mathbf{x})). \qquad (1)$$

However, most of known methods were developed for classification problems. It is generally recommended to use the arithmetic average of predictions of base models, for regression analyses purposes. Such function is, however, sensitive to possible appearance of outliers, the presence of which may degrade the accuracy of prediction.

One of the possibilities to reduce the impact of outliers is to use the weighted average in the similar way as for classification models, where the three classes of aggregation methods are the most commonly used (Gatnar, 2008):

- majority voting, the model assigns an observation to the class chosen by the largest number of base models:

$$\hat{D}^*(\mathbf{x}_i) = \arg\max_j \sum_{m=1}^{M} I(\hat{D}_m(\mathbf{x}_i) = C_j), \qquad (2)$$

[1] Corresponding author: University of Economics in Katowice, Department of Statistics, 1-go Maja 50, 40-287 Katowice, Poland, e-mail: jacek.stelmach@polwax.pl.

- weighted majority vote, prediction result is defined according to the formula:

$$\hat{D}^*(\mathbf{x}_i) = \arg\max_j \sum_{m=1}^{M} w_m I(\hat{D}_m(\mathbf{x}_i) = C_j), \tag{3}$$

- using naïve Bayesian classifier with the vector of probabilities *a posteriori*:

$$\hat{D}^*(\mathbf{x}_i) = \arg\max_j \{W_j(\mathbf{x}_i)\} \tag{4}$$

where $I$ – indicator function, $C_j$ – class indicated by base model for [$\mathbf{x}_i, y_i$] observation, $d_m$ – prediction result, $W_j$ – discrimination index:

$$W_j(\mathbf{x}_i) = p(C_j) p(d_1,..., d_M \mid C_j). \tag{5}$$

This paper presents the results of the experiments that implements second method for regression purposes:

$$\hat{D}^*(\mathbf{x}_i) = \frac{1}{\sum_{m=1}^{M} w_m} \sum_{m=1}^{M} w_m \hat{D}(\mathbf{x}_i) \tag{6}$$

where $w_m \geq 0, m = 1, 2, \ldots, M$ and $\sum_{m=1}^{M} w_m \neq 0$.

The crucial problem is how to determine the weights to get the most accurate aggregated model. It should be a compromise between fitting the weights to learning sample and ability to generalize a modeled phenomenon.

## 2. Experiment description

### 2.1 Datasets

The experiment was carried with both types of datasets: simulated and empirical.

Simulated datasets include (200 observations created):

*Dataset 1*. Uncorrelated predictors $x_1,..., x_{10}$ ~ N(0,1), dependent variable according to formula:

$$y = \sum_{i=1}^{10} \beta_i x_i + \frac{e}{10} \tag{7}$$

where $\beta_1,..., \beta_{10}$ - the coefficients sampled from range (-1, 1), $e$ ~ N(0,1) – random error.

*Dataset 2*. Uncorrelated predictors $x_1,..., x_{10}$ ~ N(0,1), dependent variable according to formula:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 x_4 + \frac{e}{10} \tag{8}$$

where $\beta_i$, $e$ – as above.

*Dataset 3.* Predictors $x_1,...,x_{10}$ correlated according to Table 1 with zero means vector, dependent variable according to formula (7).

*Dataset 4.* Predictors $x_1,...,x_{10}$ correlated according to Table 1 with zero means vector, dependent variable according to formula (8).

*Dataset 5.* Predictors $x_1,...,x_{10}$ correlated according to Table 2 (higher dependencies) with zero means vector, dependent variable according to formula (7). Additionally for 5% of random observations, the value of dependent variable was replaced by the value five times bigger (simulation of additional distortion).

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1.0 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0.4 | 1.0 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.4 | 1.0 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.4 | 1.0 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.4 | 1.0 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 1.0 | 0.4 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 1.0 | 0.4 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 1.0 | 0.4 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 1.0 | 0.4 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 1.0 |

**Table 1.** Covariance matrix of *"Dataset 3"* and *"Dataset 4"*.

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1.0 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.9 |
| 0.7 | 1.0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.9 |
| 0.7 | 0.5 | 1.0 | 05 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.6 |
| 0.7 | 0.5 | 0.5 | 1.0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.6 | 0.7 |
| 0.7 | 0.5 | 0.5 | 0.5 | 1.0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.6 |
| 0.7 | 0.5 | 0.5 | 0.5 | 0.5 | 1.0 | 0.5 | 0.5 | 0.5 | 0.6 |
| 0.7 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1.0 | 0.5 | 0.5 | 0.6 |
| 0.7 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1.0 | 0.5 | 0.6 |
| 0.7 | 0.5 | 0.5 | 0.6 | 0.5 | 0.5 | 0.5 | 0.5 | 1.0 | 0.6 |
| 0.9 | 0.9 | 0.6 | 0.7 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 1.0 |

**Table 2.** Covariance matrix of *"Dataset 5"* and *"Dataset 6"*.

*Dataset 6.* Predictors $x_1,...,x_{10}$ correlated according to Table 2 (higher dependencies) with zero means vector, dependent variable according to formula (8). Additionally for 5% of random observations, the value of dependent variable was replaced by the value five times bigger (simulation of additional distortion).

Three next simulated datasets coming from Friedman (1991) proposal. Friedman recommends presented generators to create a multi-dimensional samples that put high demands on non-parametric regression methods due to their non-linearity and random components.

*Dataset 7.* Uncorrelated predictors $x_1,...,x_{10}$ ~ N(0,1), dependent variable according to formula:

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + e \tag{9}$$

where $e$ ~ N(0,1) – random error.

*Dataset 8.* Four predictors uniformly distributed in the ranges:

$$0 \le x_1 \le 100$$
$$40\pi \le x2 \le 560\pi$$
$$0 \le x_3 \le 1$$
$$0 \le x_1 \le 100$$

dependent variable according to formula:

$$y = \left[ x_1{}^2 + \left( x_2 x_3 - \frac{1}{x_2 x_4} \right)^2 \right]^{0.5} + e \tag{10}$$

where $e$ ~ N(0,9) – random error.

*Dataset 9.* Four predictors uniformly distributed as for *Dataset 8*, dependent variable according to formula:

$$y = \tan^{-1}\left( \frac{x_2 x_3 - \frac{1}{x_2 x_4}}{x_1} \right) + e \tag{11}$$

where $e$ ~ N(0,1) – random error.

Empirical datasets were chosen from *UCI Machine Learning Repository* as in Table 3.

The variables: *PRP* in *„Computer”* dataset, *str* in *“Concrete”* dataset, *medv* in *“Housing”* dataset and *pw* in *“Iris”* dataset were set as dependent variable, all the rest of variables was set as predictors. Because *“Iris”* dataset contains three classes: S*etosa*, *Versicolor* and *Virginica*, the experiment was carried out separately for each class.

| Dataset name | Observations number | Variable number | Repository address |
|---|---|---|---|
| *Computer* | 209 | 7 | http://archive.ics.uci.edu/ml/datasets/Computer+Hardware |
| *Concrete* | 1030 | 9 | http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength |
| *Housing* | 506 | 14 | http://archive.ics.uci.edu/ml/datasets/Housing |
| *Iris* | 150 | 4 | http://archive.ics.uci.edu/ml/datasets/Iris |

**Table 3.** A list of empirical datasets used in the experiment.

## 2.2 Methods of weights selection coming from classification models

Very important issue that determines the quality of aggregated models should take into account the minimization of forecasts outliers while preserving the basic advantage of aggregation: reducing the aggregate forecast error. Therefore there were proposed four methods, main idea of which comes from the classification models:

**W1.** All the weights are equal (standard arithmetic average):

$$w_m \sim \frac{1}{M} \quad , \tag{12}$$

**W2.** The weights depend on the *MAPE$_m$ (Mean Absolute Percentage Error)* of *m*-th base models prediction:

$$w_m \sim \frac{1}{MAPE_m} \quad , \tag{13}$$

**W3.** The weights, in a way proposed for classification models by Littlestone and Warmuth (1994) are proportional to the expression below, where *SE$_m$* is standard error of *m*-th base models prediction:

$$w_m \sim \frac{\max(SE_m)}{SE_m}\left(1 - \frac{SE_m}{\max(SE_m)}\right) = \frac{\max(SE_m)}{SE_m} - 1, \tag{14}$$

**W4.** The weights depend on *SE$_m$:*

$$w_m \sim \frac{1}{SE_m}. \tag{15}$$

## 2.3 Methods of weights selection with genetic algorithms

The leading idea of methods presented below was to find such vector of weights which minimizes the error of aggregated prediction. It needs to examine all the possible combinations of weights for each base model. According to Opitz and Maclin (1999), Breiman (1996) at least 25 base models is an optimal number for classification purposes. Stelmach (2013) verified that similar number of base models is necessary for regression models. In this experiment $m$ = 30 base models were chosen. It means that the inspection of all combinations may require very high, unacceptable time. Therefore it was proposed to use genetic algorithms for weights selection purposes. Of course as heuristic method, it does not ensure the optimal solution, but most often the result is close to such a solution.

Genetic algorithms are class of the evolutionary algorithms invented by Holland (1975) in 1970s, to find the solutions to optimization and search problems. It is the method that uses evolutionary approach inspired by the principles of biological evolution, coding a number of parameters into strings or chromosomes (genotypes) to generate another solution approaching the optimum (Goldberg, 1989). The algorithm presented in figure 1 includes (Mitchell, 1998):

- **selection** – basing on chosen selection operator (fitness function), only genotypes with the best fitness value can be reproduced,
- **crossover** – forms new offspring from two parent genotypes by combining part of the genes from each, randomly choosing the point (points) of crossover and the number of such points,
- **mutation** – it randomly alters the value of genes with certain probability, recommended (based on real evolution process) is to mutate few genes.

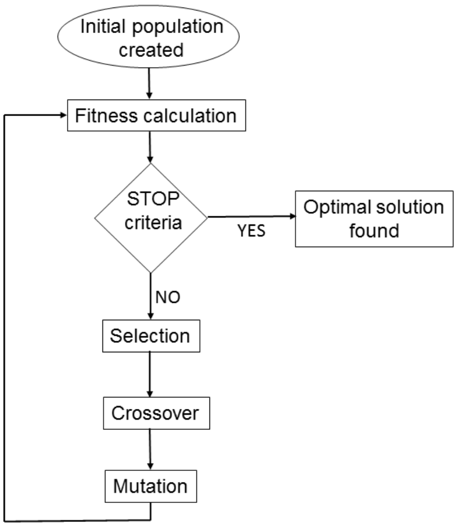


**Fig. 1.** Genetic algorithm.

Important issue in genetic algorithm is a choice of fitness function, its value decides which genotypes can create another generation and influences the final results. In the experiment each weight associated with each base model forecast was coded on 5 bits of genotype with *SE* (standard error of aggregated model prediction) as fitness function. Such coding in the most obvious case allows choice of the weights in the range <0, 32>. However, it may cause too drastically reduce the magnitude of predictions of some base models. Therefore four cases of range of weights were assumed: <0, 32> (**GA1**), <1, 32> (**GA2**), <10, 42> (**GA3**), <32, 64> (**GA4**).

### 2.4 Monte Carlo simulation

The base models, using the above datasets were created with three types of methods:

1. *bagging* (bootstrap aggregating) sampling with OLS method,
2. *bagging* sampling with neural network (MLP type) method,
3. *bagging* sampling with regression trees method.

The datasets were 100 times randomly divided into two parts: validation set (10 observations) and training (learning) set – all other observations. Based on training set the weights calculations were carried out to obtain aggregated forecasts. All the weights, calculated according to method presented above were standardized (sum of all the weights has to be equal to one). As an indicators of quality of aggregated models two measures were calculated:

- *RSE – Residual Square Error* for training set,
- *MAE – Mean Absolute Error* for validation set.

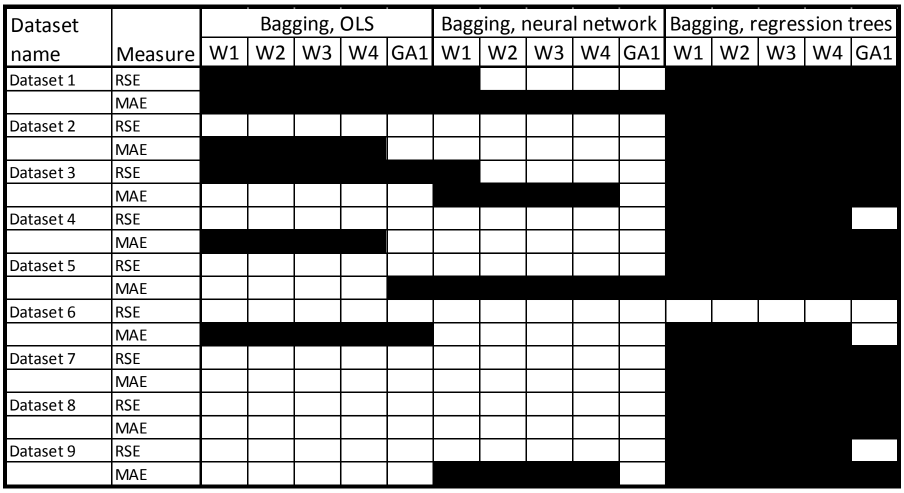
The same measures were calculated for "the reference model" obtained by OLS method.

Both measures were the basis for assessing the impact of the methods of choosing the weights on the quality of the obtained regression models.

## 3. Experiment results

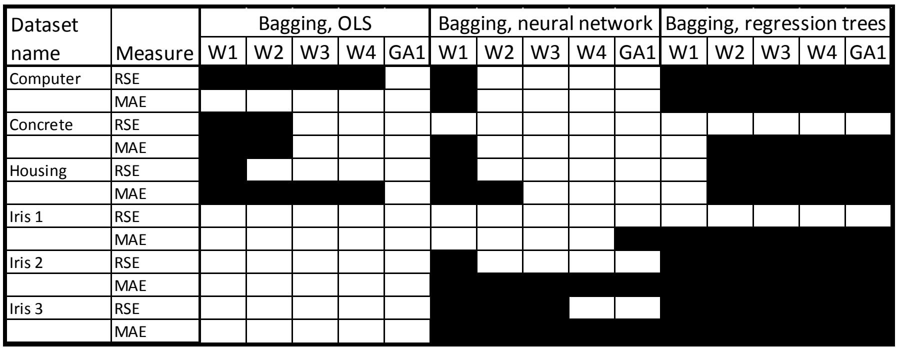
Regardless of the range of the weights values obtained with genetic algorithm (**GA1** – **GA4**), calculated measures were very close. Therefore the discussion of the results was limited to the case **GA1**. Because of huge number of data, the results of the experiment is presented in Figure 2 (simulated datasets) and Figure 3 (empirical datasets).

Black color marks a case in which a value of *RSE* or *MAE* measure for particular method is higher than the value for OLS method. Table 4 shows the value of measures for empirical datasets that shows the differences of *MAE* prediction error for OLS model and aggregated model created with *bagging* – OLS.



| Dataset name | Measure | Bagging, OLS | | | | | Bagging, neural network | | | | | Bagging, regression trees | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | W1 | W2 | W3 | W4 | GA1 | W1 | W2 | W3 | W4 | GA1 | W1 | W2 | W3 | W4 | GA1 |
| Dataset 1 | RSE | ■ | ■ | ■ | ■ | ■ | ■ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| | MAE | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Dataset 2 | RSE | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| | MAE | ■ | ■ | ■ | ■ | □ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| Dataset 3 | RSE | ■ | ■ | ■ | ■ | ■ | ■ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| | MAE | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | □ | ■ | ■ | ■ | ■ | ■ |
| Dataset 4 | RSE | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | □ |
| | MAE | ■ | ■ | ■ | ■ | □ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| Dataset 5 | RSE | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| | MAE | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Dataset 6 | RSE | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ |
| | MAE | ■ | ■ | ■ | ■ | ■ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | □ |
| Dataset 7 | RSE | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| | MAE | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| Dataset 8 | RSE | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| | MAE | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| Dataset 9 | RSE | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | □ |
| | MAE | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | □ | ■ | ■ | ■ | ■ | ■ |

**Fig. 2.** Coded results obtained for simulated datasets.



| Dataset name | Measure | Bagging, OLS | | | | | Bagging, neural network | | | | | Bagging, regression trees | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | W1 | W2 | W3 | W4 | GA1 | W1 | W2 | W3 | W4 | GA1 | W1 | W2 | W3 | W4 | GA1 |
| Computer | RSE | ■ | ■ | ■ | ■ | □ | ■ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| | MAE | □ | □ | □ | □ | □ | ■ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| Concrete | RSE | ■ | ■ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ |
| | MAE | ■ | ■ | □ | □ | □ | ■ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ |
| Housing | RSE | ■ | □ | □ | □ | □ | ■ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ |
| | MAE | ■ | ■ | ■ | ■ | □ | ■ | ■ | □ | □ | □ | □ | ■ | ■ | ■ | ■ |
| Iris 1 | RSE | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ | □ |
| | MAE | □ | □ | □ | □ | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ | ■ |
| Iris 2 | RSE | □ | □ | □ | □ | □ | ■ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| | MAE | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| Iris 3 | RSE | □ | □ | □ | □ | □ | ■ | ■ | ■ | □ | □ | ■ | ■ | ■ | ■ | ■ |
| | MAE | □ | □ | □ | □ | □ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |

**Fig. 3.** Coded results obtained for empirical datasets.

| Dataset | *MAE* values | | | | | |
|---|---|---|---|---|---|---|
| | **OLS** | **W1** | **W2** | **W3** | **W4** | **GA1** |
| *Computer* | 42.34 | 41.71 | 40.60 | 40.60 | 41.94 | 40.70 |
| *Concrete* | 8.76 | 8.77 | 8.77 | 8.75 | 8.75 | 8.69 |
| *Housing* | 3.07 | 3.08 | 3.08 | 3.08 | 3.07 | 3.01 |
| *Iris 1* | 0.837 | 0.831 | 0.831 | 0.831 | 0.831 | 0.830 |
| *Iris 2* | 0.089 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 |
| *Iris 3* | 0.223 | 0.217 | 0.217 | 0.217 | 0.217 | 0.211 |

**Table 4.** *MAE* values of prediction for models created with OLS method and aggregation methods based on OLS base models.

## Conclusion

Aggregated regression models allow to reduce forecast error in most of the analyzed cases, comparing to Ordinary Least Square method. Additionally, the use of a weighted average as an aggregating function reduced this error. There were no significant differences resulting from the different methods of calculating the weights although in most cases the use of the genetic algorithm allowed to obtain the most accurate forecasts. During creating base models the method of creation must be carefully chosen. In presented results, the highest value of errors (even higher than for OLS models) was observed for base models obtained with regression trees method.

## Acknowledgements

The work paper funded by resource on science period 2012-2015, National Science Centre No. DEC-2011/03/B/HS-4/05630.

## References

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *26*(2), 123-140.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *Annals of Statistics*, *19*(1), 1-67.

Gatnar, E. (2008). *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji.* Warszawa: Wydawnictwo Naukowe PWN.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems.* Ann Arbor: The University of Michigan Press.

Littlestone, N., & Warmuth, M. (1994). Weighted majority algorithm. *Information and Computation*, *108*, 212-261.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169-198.

Stelmach, J. (2013). On estimation of a quantity of base models with parametric and permutation tests. *Acta Universitatis Lodziensis, Folia Oeconomica*, *286*, 79-86.