Janusz Dąbrowski*, Tomasz Adamczyk**

Application of GAM Additive Non-Linear Models to Estimate Real Estate Market Value

1. Introduction

Generalized additive models (Generalized Additive Models – GAM) have a relatively short history, and hence their use is difficult. Although Hastie and Tibshirani [1] and Schimek [3] provide a detailed interpretation of them, yet in professional practice these methods were not used. In the literature for valuers there was not found a single citation for the application of GAM models. The most characteristic for generalized additive models of data analysis is a graph of the expected values and sub-rests, thanks to it we can see the final match of cubic (third degree curve) concatenated to the final model. Residual statistics as in other methods allow us to assess the quality of the fit and predict stand-off values.

Developers often use non-linear estimation by calculating the relationship between the price of a property and the time needed to sell it. A linear relationship is assumed a priori in multiple regression analysis of variance. Nonlinear estimation allows the optimal choice of function which binds variables and appoints parameters. This may be, among others, a function: logarithmic, exponential, or being the product of independent variables. Linear function is relatively simple to interpret, in contrast to the nonlinear function where predictors are smoothed by the function of the third degree, which allows for a better fit of the model to follow. Hastie and Tibshirani (1990) suggest that *R*-square value, calculated as a relative improvement of the overall variance for the final model, was considered as a measure of fit of the model.

To the presented subject there can be formulated the following argument:

Non-parametric GAM models has allowed an extensive analysis of the real estate market, together with an indication of the boundary conditions for the analyzed parameters of the wider housing market environment.

* The Bronisław Markiewicz State School of Higher Vocational Education, Jarosław
** Faculty of Mining Surveying and Environmental Engineering, AGH University of Science and Technology, Krakow

In order to develop the thesis, the author of the publication presents the process of GAM statistical models built on global parameters selected on the basis of factor analysis.

2. General Characteristics of Non-Parametric Models GAM

Generalized additive models (GAM) can be built for different distributions (normal, gamma, and Poisson) and for various binding functions. The Statistica program generates a full range of statistics of a model (including the predicted and observed values) and cubic functions of folding charts. The procedure GAM logit model enables the modeling of variable-binding function of the binomial logit. By identifying the number of degrees of freedom, we can get the smoothness of each predictor of quantification. Low number of degrees of freedom results in worse fit, however it gets a cubic function close to a linear function. The model can be built with or without an intercept with the exception of a model built on the basis of quality parameters which is always required where there is an intercept. In the Satistica program generalized additive models are a generalization of multiple regression.

The differences between the generalized linear model and the linear model:

the distribution of the dependent variable can be different from the normal distribution and it doesn't have to be the continuous distribution, the values of the dependent variable are predicted on the basis of a linear combination of predictors, that are "associated" with the dependent variable by the link function.

In linear regression most often by the use of the least squares method we fit in the line hyperplane for the set of predictors X in order to carry out prediction Y (of the dependent variable) which can be shown with the formula:

$$Y = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + a_3 \cdot X_3 + \ldots + a_m \cdot X_m \tag{1}$$

where:

Y – the value of the predicted dependent variable,
$X_1 \ldots X_m$ – values of predictors,
$a_0, \ldots, a_m$ – regression coefficients.

The generalized model of multiple regression is received by retaining the additive nature of the model with the change of components of the linear equation $a_i \cdot X_i$ into $f_i(X_i)$ where f_i are non-parametric functions of the predictor X_i. In the additive models method we estimate the function of each of the predictors, thanks to which we receive better prediction of the dependent variable.

In practice we estimate the (non-parametric) function of each of the predictors, thanks to which we receive a better estimation of the dependent variable value. The relations of linking the variable *Y* with the values of the variable *X* in the generalized linear model can be shown with the following function:

$$Y = g(a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + a_3 \cdot X_3 + \ldots + a_m \cdot X_m) \tag{2}$$

The function inverse to function *g*(...)can be shown by the following formula:

$$g_i(Y_{mi}) = a_0 + a_1 \cdot X_1 + a_2 \cdot X_2 + a_3 \cdot X_3 + \ldots + a_m \cdot X_m \tag{3}$$

where Y_{mi} – signifies the expected value *Y*.

We have the following most often used link functions:

link function – Log: $f(z) = \log(z)$ (4)

link function – reciprocal: $f(z) = 1/z$ (5)

identity link function: $f(z) = z$ (6)

The aim of the generalized additive models is optimization of dependent variable *Y* prediction by estimation of unknown (non-parametric) functions of predictors, which are linked with the dependent variable with the help of the binding function. Finally instead of estimating parameters (e.g. significance of regression in multiple regression), in the generalized additive models we calculate the general unknown function (non-parametric) linking expected values *Y* (after the transformation) with the predictor values.

Generalized linear models also have disadvantages. On one hand they are very flexible and they give good matching in the case of non-linear dependences and great "unclearness" of predictors, on the other hand the excessive matching can cause a decline in prediction credibility. Therefore one should be extremely cautious building the model, so as not to use this flexibility to excessively match to data, i.e. not to create the too complex model (with many degrees of freedom), which provides good matching, not appearing however in successive control research. One should also compare the quality of matching received as a result of the analysis in generalized additive models with matching received in generalized linear and non-linear models. In other words – one should estimate whether the additional complexity (generality) of generalized linear models (regressive matching) is necessary to receive better matching. It is often not the case, and in the case of many comparable models the simple model is the better one. These issues are described in detail in Hastie and Tibshirani (1990). It is definitely easier to interpret results received in (generalized) linear models in comparison to generalized additive models, which is a very important element in making decisions concerning the choice of the model. Sometimes it is better to build a model which is less credible and simpler in results interpretation than a model with great matching but abstract concerning results interpretation. Linear models are much easier to interpret and to predict especially for people who do not work with statistics regularly. The difficulties of interpretation are the main disadvantage of generalized additive models, whereas their basic advantage is better matching and prediction quality.

The methods available in generalized additive models of the Statistica program are the implementation of techniques developed and popularized by Hastie and Tibshirani [1]. The cubic function is splined with smoothing technique used in scatter diagrams 2W (see description of Statistica program), which provides smoothed form of dependence between two variables. The cubic spline function is often used in the generalized additive model in order to find a non-parametric function of predictors which gives the most accurate predictions of the dependent variable value. The detailed information about smoothing by the cubic spline function and comparison with other methods of smoothing is available in [1, 3].

3. Non-Linear Additive Models – GAM in the Real Estate Market Analysis

In this work we analyzed the results of 73 models and best models are summarized in table 1

While analyzing the property market one can very efficiently use the generalized models of the GAM to analyze the real estate market and the prediction of the market value of the property. The article presents the final results of the research work carried out on attractive real estate markets (Krakow, Warsaw) in the years 1999–2007 in the context of the analysis of the real estate market environment and the impact of environment on property prices. During the period 205 different types of socio-economic indicators of macro-economic nature were analyzed. The results underwent factor analysis, which resulted in reduction of independent variables. As a result, there were finally 73 models selected. Each model consisted of a variety of global parameters (from one to seven) to describe the real estate market environment. The statistical analysis was performed by: the study of residuals, the coefficients *R*2, reliability of coefficients and analysis of statistical graphics. The basic research problem was to get the statistical significance of coefficients of non-linear additive models.

Table 1. Statistical models – GAM

Date in months	Statistical models – GAM				
	321	501	101–103	Time-GAM	Time
	The rest for selected models, in zlotys				
30	–16	38	–71	–97	33
29	208	273	129	122	240
28	–49	13	–30	–18	87
27	–41	–160	–202	–137	–46
26	158	97	198	176	251
25	–155	–253	94	–89	–32
23	–70	–3	58	70	81
22	48	66	50	105	88
21	–221	–91	30	–78	–125
20	–17	–3	–86	–9	–88
19	187	129	–60	111	–1
18	181	155	117	130	–14
15	–199	–215	–232	–240	–443
14	–262	–231	–186	–250	–446
13	–60	–117	–64	–126	–299
12	66	30	116	15	–122
11	48	24	65	16	–75
9	130	157	234	173	179
8	124	105	109	127	176
7	2	28	–41	48	133
5	62	49	–163	43	176
2	–80	–64	–86	–38	127
0	–45	–29	21	–54	121
Minimum	–262	–253	–232	–250	–446
Maximum	208	273	234	176	251
Median	–16	24	21	15	33
Standard deviation	132	133	127	121	193
The average absolute value of residuals	106	101	106	99	147
Sum (residuals) 2/1000	384	387	356	324	818
The average absolute value of residuals as a percentage	72%	69%	72%	67%	100%
Sum (residuals) 2/1000 in %	47%	47%	44%	40%	100%

Signs of symbols used in the model:

124 – inflation y/y (Euro), in %
130 – retail sales m/m (Euro) incrementally, in %
134 – underlying inflation y/y (U.S.), in %
202 – boom in the construction increasing, in points
209 – sell-off ratio, in %
217 – rate expectations, in days
321 – index divisia M3
326 – loan refinanced, in mln zlotys
410 – IPSOS poll (of a constant price)

Table 1 summarizes the residues and basic statistics for the four basic models and model comparison. These are models that have successfully passed the statistical verification. In calculating the rest of the models included in tables 1 and 2 it can be stated that much of it meets very well the first and second test of the value of obtaining adequate rest, together with high *R*-squared materiality. Only a few models, a small number of independent variables have successfully verified statistically significant regression of coefficients which can prove that in the GAM models more simple models are preferred. Table 2 presents summary statistics.

Table 2. Summary statistics for the non-linear additive models

Model	Summary statistics for the GAM models						
	final	rest	number	external	number	rating	*R*-squared
321	383683.6	17.99639	23	1	3	21320.03	98.17
501	386551.3	17.99793	23	1	3	21477.55	98.16
101–103	356025.6	13.99655	23	1	15	25436.66	98.30
Time GAM	323972.0	18.00216	23	1	3	17996.29	98.45

In the table 2 we observe a very high *R*-squared value. These results confirm the thesis of a good match of model to observed values. The result of coefficient correlation of above 0.9 is quite "common" for a substantial part (several) models tested. The most difficult requirement to meet the statistical method in the GAM was to obtain statistically significant predictor of the coefficients. In practice, for each predictor it is created a worksheet in which there are values of the original predictor and smooth match (with 95% confidence interval for the values match), and sub-rests treated as if the model is not affected by any other predictors in question). This last phase of testing the model is very difficult and here there are rejected a lot of models that have passed the first two stages of screening. The following table summarizes the coefficients and their statistics. The value of likelihood of *p*-non-linear is the main criterion for allowing a model for further practical use. In addition, all tested models also have the correct value of the ratio of the GAM coefficients to the error.

The framing of the time variable results from the presumption formed above depending on treatment of the smaller GAM models as a peculiar type of computational algorithm aiming at obtaining smaller residual value. In the table the distinction of the GAM model concerning time flow from the statistics of two-dimensional linear regression the variable was marked as GAM Time. A very superficial analysis shows that the model GAM Time is the best among the analyzed ones.

Table 3. Summary statistics for the GAM models

Number of global attributes		Summary statistics for the GAM models					
		index	degrees	GAM	error	standardized	prob.
1	intercept	0	1.0000	–2830.65	306.439	–9.237	
	321 index divisia M3	1	4.0036	31.53	1.026	30.745	0.0006
2	intercept	0	1.0000	573.554	202.792	2.828	
	501 number of immigrants, in millions	1	4.0021	3627.074	121.783	29.783	0.0000
3	intercept	0	1.0000	9737.154	534.253	18.226	
	101 GDP q/q, in %	1	4.0010	280.105	39.920	7.017	0.0075
	103 the unemployment rate, in %	2	4.0025	–308.051	22.085	–13.948	0.0000
4	intercept	0	1.0000	–2085.84	259.487	–8.038	
	time GAM	1	3.9978	106.95	3.197	33.454	0.0000

Such a thesis is justified by the values of the standard deviation and the average value of the absolute value of residual. Using the GAM model for time increased the accuracy of the model's estimation about 30–50% of the value of particular statistics. The final verification of the model GAM Time is the illustration of residual and their absolute values. The residual value and their absolute values illustrated in figures 1 and 2 confirm the thesis concerning good matching of the GAM Time model to observed values. Using the model remarkably improved the accuracy of prediction.

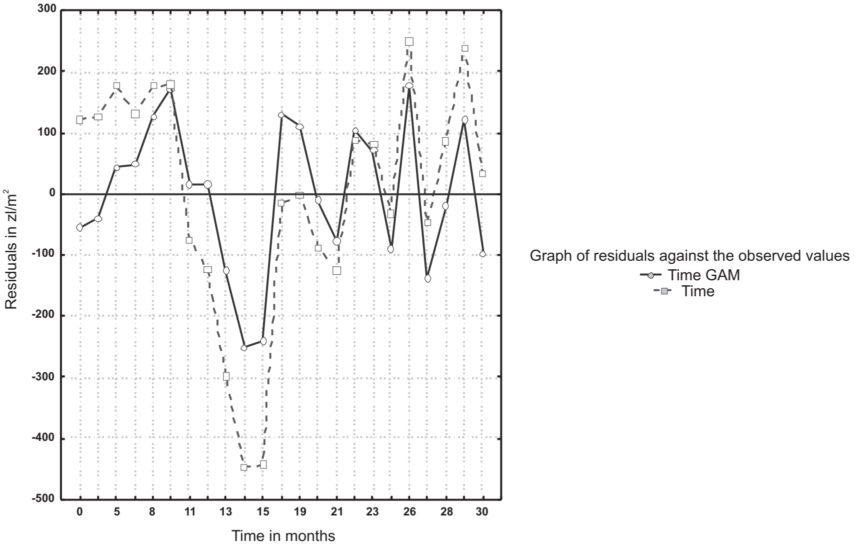

Fig. 1. The diagram of residual dispersion

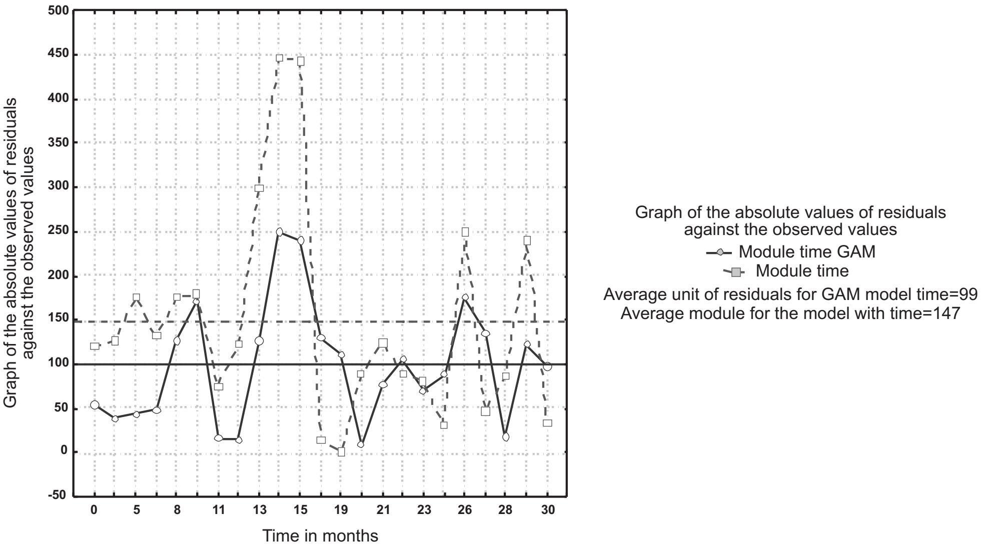

Fig. 2. The dispersion diagram of absolute values of residual

4. Summary

Conducting the analysis of 73 statistical models confirmed the legitimacy of using the GAM method only for models with a small number of global parameters. After exceeding two parameters none of the models positively underwent four-level statistical verification. Surprisingly good matching was achieved using the GAM method for time. The sum of residua squares and the average absolute value of residual was respectively 40 and 67 percent compared values for time (of the model of two variables regression). Graphic illustration of residual, especially absolute values fully confirmed the model's good matching and the high quality of prediction. The research shows that using GAM models provides good results in real estate market analysis and it can be successfully used to predict the real estate value.

References

[1] Hastie T.J., Tibshirani R.J.: *Generalized Additive Models*. Chapman & Hall, New York 1990.

[2] Schimek M.G.: *Smoothing and regression: Approaches, computations, and application*. Wiley, New York 2000.

[3] McCullagh P., Nelder J.A. *Generalized linear models*. 2nd Ed., Chapman & Hall, New York 1989.