EKONOMETRIA ECONOMETRICS 32 • 2011

ISSN 1507-3866

Ewa Szabela-Pasierbińska
Wrocław University of Economics

THE ROLE OF DATA AGGREGATION IN THE PROCESS OF SELECTING A SALES FORECASTING METHOD

Summary: The paper presents the role of aggregation in the process of sales forecasting. While explaining the term "aggregation" it was noticed that the notion appears in forecasting at the stage of forecast data analysis and at the stage of forecast construction. In the former case aggregation, although it might cause the loss of certain amount of information on the analysed phenomenon, facilitates or even enables selecting a forecasting model indispensable for establishing a forecast for this phenomenon. In the latter case aggregation enables obtaining forecasts for longer periods or larger areas than the initially established forecasts. The theoretical discussion was illustrated with an empirical example related to forecasting sales of household goods and appliances.

Key words: data aggregation, forecasting stages, forecast, forecast disaggregation.

1. Introduction

The selection of the forecasting method is a consequence of the previously specified forecasting task in which the identification and description of the forecast issue and gathering of appropriate data are performed. The choice leads directly to constructing a forecasting model and building a forecast. The forecasting model constructed in the selected method is supposed to describe regularities found during data analysis in the analysed phenomenon or between this phenomenon and other phenomena. It is frequently the case, however, that good data in terms of quality do not display regularities which could be described with the use of a formal mathematical model in a satisfactory manner.

The aim of the paper is to select an appropriate forecasting model to enable determining a forecast of the sales volume of a given household appliance. The statement that data aggregation is indispensable in certain situations for constructing a formal forecasting model and subsequently for establishing a good quantity forecast on its basis was assumed as the null hypothesis.

2. The term "aggregation"

Aggregation (in Latin *aggregatio*) denotes a process of joining elements together to form one whole and also the state of aggregation of elements. The result of aggregation is an aggregate, that is a whole obtained through joining of heterogeneous elements [*Słownik 100 tysięcy*... 2005, p. 6]. A broader definition of aggregation is provided by Tadeusz Bołt, who states that it is "a process of joining individual units of a population, and subsequent transformation of the realisation value of each distinguished feature of individual units into one value referring to the elements' population". Bołt adds that the need for aggregation always occurs when the taken decisions refer not to individual units but to groups of the units which constitute a statistical population. He also mentions the issue of aggregation consistency, which will occur when the results of the conducted analysis of the problem based on aggregated data are identical with those obtained after the use of non-aggregated data [Bołt et al. 1985, p. 16-17]. It is worth quoting also the definition by John Green, which focuses more on the economic and numerical dimension (nature). According to Green, aggregation is a process due to which a portion of the available information necessary to solve a problem is sacrificed in order to simplify the solution of the problem […]. Aggregation acquires the form of replacing a set of numbers (e.g. the number or prices of goods) with one number or a smaller set of numbers or *aggregates*. Green points to the fact that aggregation ought to be employed every time the cost of processing more detailed data in the examination is higher than the reliability of the results of the examination [Green 1964, p. 3].

Where statistical data in the form of one or multidimensional time series, one or multidimensional profile series, or profile and time series are used to describe how a given phenomenon is shaped, aggregation may be identified with summing.

According to the concept, aggregation may be considered in the substantial, spatial and time aspect [Dittmann 2000, p. 43].

Substantive aggregation is the case where the data describing the examined phenomenon are summed in the objects. Spatial aggregation refers to summing data from smaller geographic areas into data from larger areas. Time aggregation occurs when the data (but only those of stream nature) of high measurement frequency are summed into data of lower measurement frequency. It needs to be noticed at this point that the aggregation process will always result in data reduction. The reduction may concern the data dimension (e.g. four-dimensional time series illustrating the sale of a product in four sales points with the use of substantive aggregation may be reduced to one-dimensional series illustrating the overall sales of the product), the number of observations in a series (e.g. a twelve-element time series illustrating a monthly sale of a product with the use of aggregation may be reduced to a series of four observations concerning a quarter sales of a product), or both at the same time.

Data aggregation, along with data transformation and integration, is one of the major tasks of knowledge management systems known as *Business Intelligence*

Systems (BI) or *Business Objects* (BO) which facilitate the decision-making process at every organisation management level [Kubiak 2005, p. 98-99]. The operation of data aggregation (reduction) is also one of the more important operations performed in OLAP technology (*On-Line Analytical Processing*) – a complex system of processing and analysing multidimensional data cooperating with BI and BO [Muryjas, Miłosz 2003, p.75].

The notion of aggregation in the theoretical sense also occurs in examinations of economic relations dealt by econometrics. The subject of the aggregation theory is transforming single relations (micro-relations) into a relation concerning certain groups of objects treated as a whole (macro-relation) Theil 1979, p. 564]. The necessity to move from a micro-relation to a macro-relation results primarily from the fact that the majority of economic theories are micro-economic in nature (they refer e.g. to a selected business entity) and the econometric estimation and hypotheses verification are based on data concerning whole groups of entities among which the occurrence of macro-economic relations is assumed. More on the subject can be found in the following works: Aigner, Goldfeld [1974], Bołt et al. [1985], Theil [1979].

3. Aggregation in the forecasting process

Forecasting is a multi-stage process. It commences with the formulation of a forecast objective, in which the phenomenon with the variable characterising it is established, as well as the quality requirements as to the forecast, the aim of the forecast, and the time to which the forecast is to refer. At this stage the forecasting procedure is often established. If the constructed forecast is to concern overall sales in an enterprise and subsequently be disaggregated into forecasts of individual product groups or even forecasts of individual products, the top-down procedure will be employed. On the other hand, if the forecaster decides to construct individual forecasts (concerning specific products or groups) and subsequently to aggregate them into a forecast of general sales in an enterprise, the bottom-up procedure will be employed (cf. [Dittmann et al. 2009, p. 18]).

At the subsequent stage the factors affecting the forecast phenomenon are established based on the knowledge as to the regularities which shape the course of the analysed phenomenon. In order that the hypotheses formulated at that stage be correct and comprehensive, it is necessary to gather and analyse forecast data, which is done at the third stage.

"The collection of valid and reliable data is one of the most time-consuming and difficult parts of forecasting […]. The forecast can be no more accurate than the data on which it is based. The most sophisticated forecasting model will fail if it is applied to unreliable data" [Hanke, Wichern 2005, p. 57]. Therefore, before the gathered data is used for creating a forecast model, its quality should be appraised by conducting above all the explorative analysis (visualisation) thereof. This will enable

detecting possible irregularities; this also permits taking a decision whether a data transformation or aggregation should be performed, which would result in the thus obtained data better describing the examined phenomenon, and as a consequence allowing to achieve a better forecast. In the case of data aggregation, time, spatial and substantive aggregation could be the case, depending on the form in which data describing the forecast phenomenon occurs and what forecast is required by the recipient.

The systematic constituent established only based on well prepared data (is a constant average level or trend, periodical fluctuations) and accidental constituent (accidental fluctuations). Thus decomposed data series reflecting the forecast variable together with the hypotheses established at the second stage of the forecasting becomes the base for the selection of the method and the construction of the formal forecast model in the fourth forecast stage.

A formal model is one which typically has the form of one or several equations and the parameters of those equations are estimated by statistical methods. A model which is good in terms of quality should reflect the regularities observed and assumed by the forecaster as well as possible, and as a consequence bring good forecasts in terms of quality.

It can also be the case of aggregation not only at the stage of data collection and analysis but also at the fifth stage of forecasting, in which the forecast is constructed (calculated). If the selected forecasting procedure is the bottom-up procedure, individual forecasts will be aggregated (summed) into an overall forecast.

The quality of thus established forecasts is proved at the subsequent forecasting stages – admissibility prior to the implementation of forecast, validity – at the moment of having a real value of the forecast variable. If the forecast fulfils the quality requirements, the employed method and forecasting procedure may be used for the further establishing of forecasts. However, if the forecast is non-admissible or invalid, the forecaster will have to verify their actions at individual forecasting stages or abandon forecasting the given phenomenon.

4. Role of aggregation in selecting the method of forecasting sales of household goods and appliances – empirical example

The demonstrated example concerns the sales of one of the household appliances manufactured in an enterprise (due to the confidentiality of information, the paper does not disclose the name of the enterprise and the "modified" forecast data is presented in a "limited" manner). The data was shared in the form of a six-dimensional time series presenting a monthly sales volume of one product in six colours over the period from January 2008 to March 2010, which for the purposes of this paper was divided into six one-dimensional time series (it was assumed that they would be variables with names colour 1, colour 2,…, colour 6). The enterprise management was interested mainly in whether and how the sales of a selected product could be

forecast, with an emphasis on the selection of a formal forecasting model. The issue was the possibility to forecast sales of both a product in general and a product by colour.

The explorative data analysis (Figure 1-6) and the establishment of standard deviations and variability coefficients of individual variables (Table 1), upon checking that the data is true to the reality (there are no systematic or random errors therein), allowed to state that all time series reveal a very great variability (considerable accidental fluctuations). In addition, the conducted Pearson linear correlation

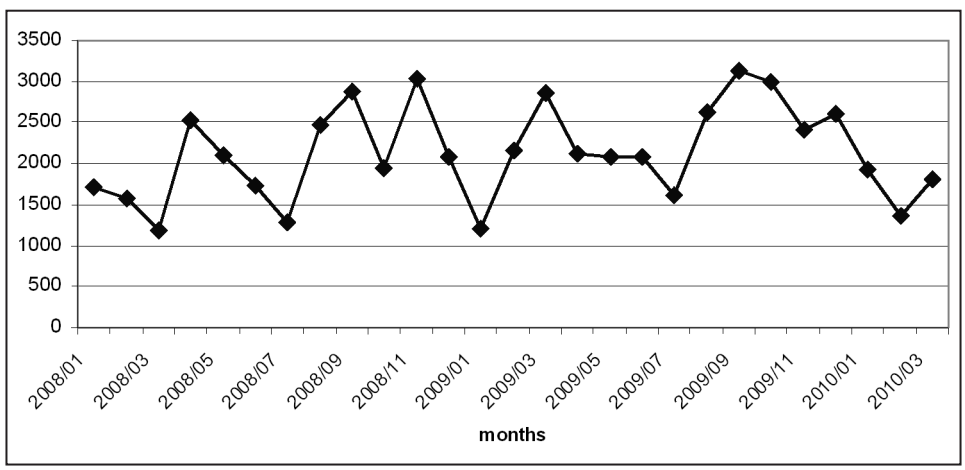

Figure 1. Monthly sales of colour 1

Source: own work based on enterprise data.

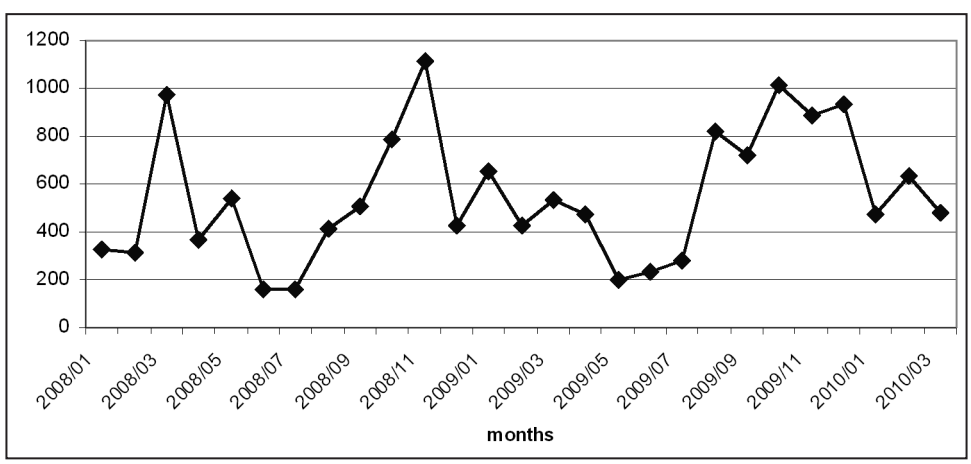

Figure 2. Montly sales of colour 2

Source: own work based on enterprise data.

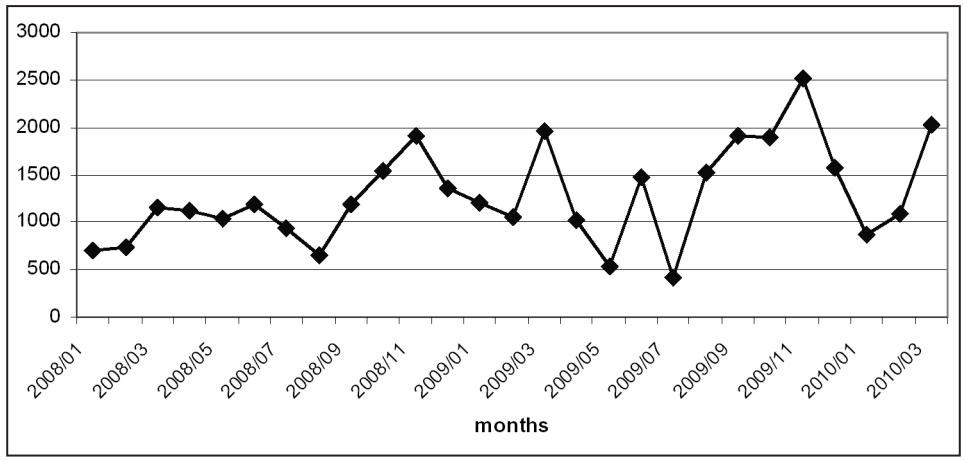

Figure 3. Monthly sales of colour 3

Source: own work based on enterprise data.

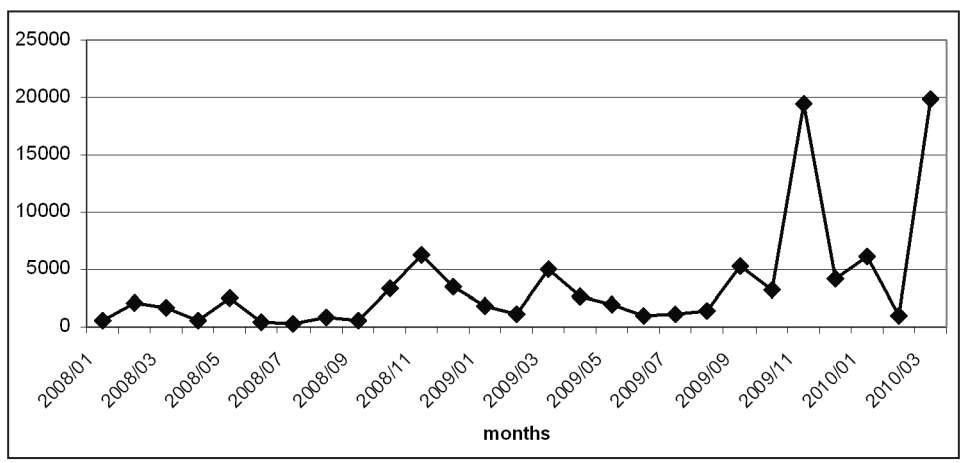

Figure 4. Monthly sales of colour 4

Source: own work based on enterprise data.

coefficient test revealed that a constant average level occurs in variables colour 1, colour 2, colour 5 and colour 6, while a slight (but significant) trend occurs in variables colour 3 and colour 4. In the case of establishing short-term forecasts, for example average-based methods should be employed in the first group of variables, and such formal models in the second group as a trend or an exponential smoothing model. Unfortunately, the great variability of all variables result in the low adaptation of the listed models to that, and possible forecasts from that being characterised by errors which are at least as great as the established (considerable) standard deviations.

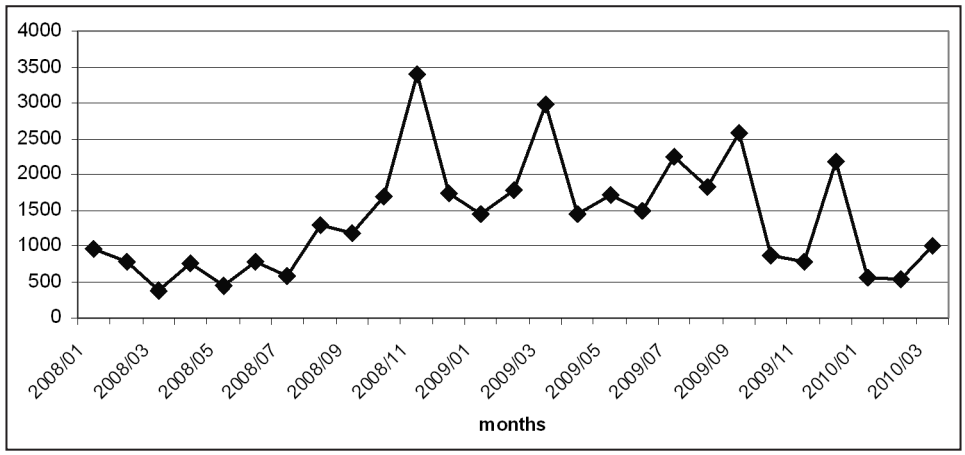

Figure 5. Monthly sales of colour 5

Source: own work based on enterprise data.

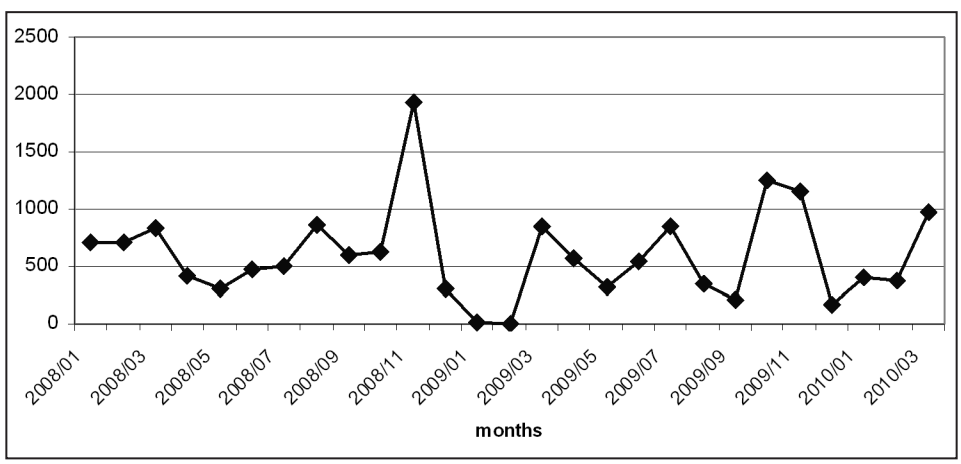

Figure 6. Monthly sales of colour 6

Source: own work based on enterprise data.

Table 1. Values of standard deviations s and variability coefficients V_s (in %) for variables colour 1,…, colour 6

Colour	1	2	3	4	5	6
s	576.1	274.3	513.5	4947.5	789.9	412.0
V_s	27.1	49.9	40.0	137.3	57.1	68.4

Source: own work based on enterprise data.

It was decided to perform time aggregation of data, in this way obtaining time series of quarterly sales volume of the product in individual colours. Those series are presented in Figures 7-12.

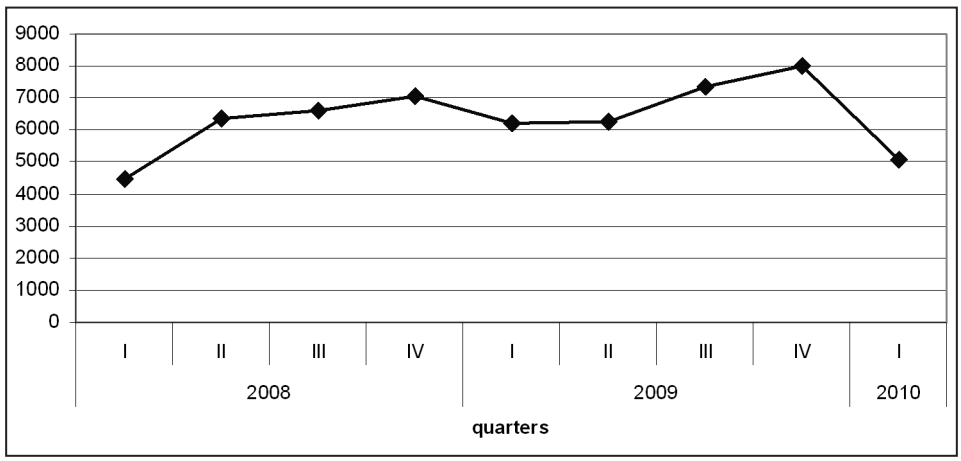

Figure 7. Quarterly sales of colour 1

Source: own work based on enterprise data.

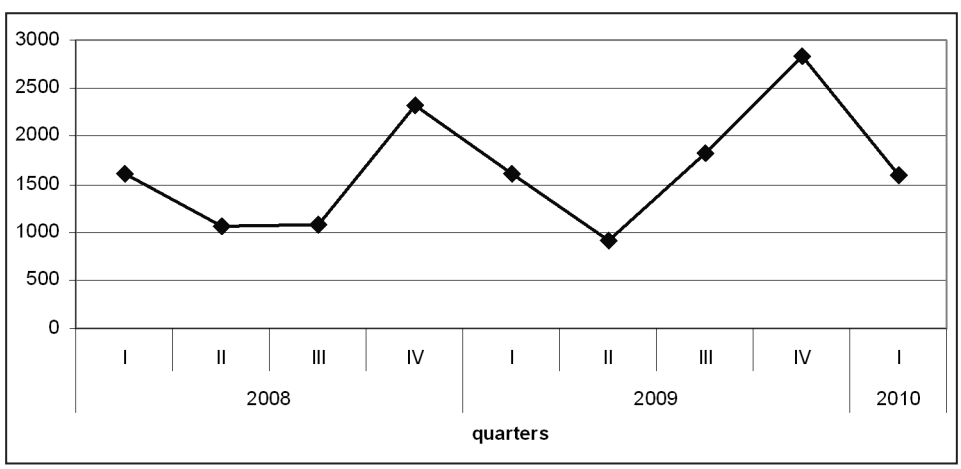

Figure 8. Quarterly sales of colour 2

Source: own work based on enterprise data.

While making a visual analysis of the figures , it can be stated that data aggregation permits identifying seasonal fluctuation basically in all quarter data – the highest sales volume is recorded almost always in each fourth quarter (except quarterly sales of product 5), which may be probably explained by the fact that in this period an

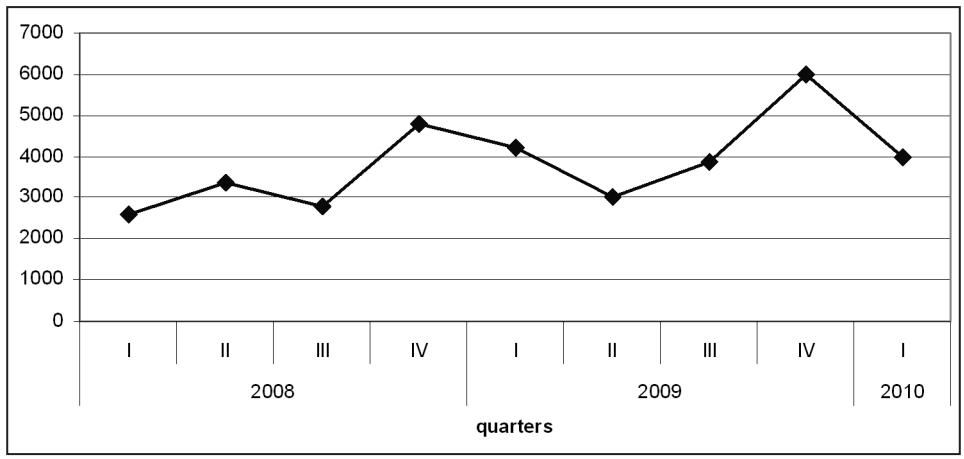

Figure 9. Quarterly sales of colour 3

Source: own work based on enterprise data.

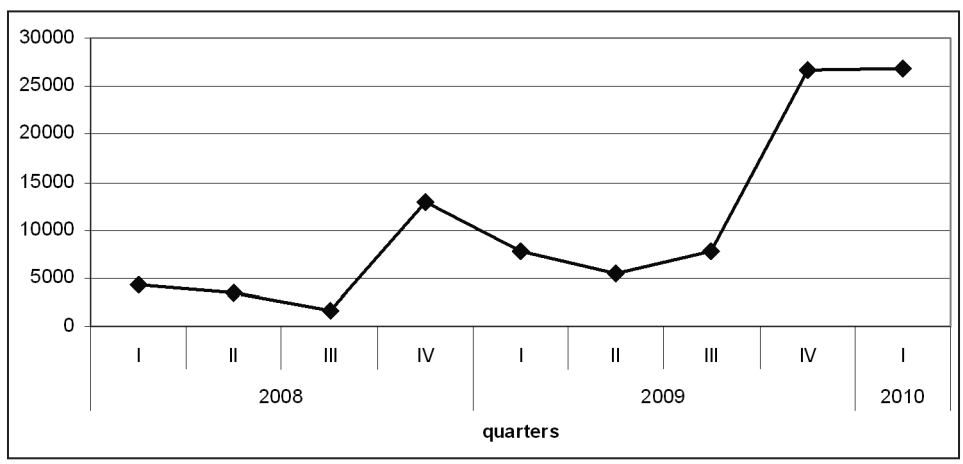

Figure 10. Quarterly sales of colour 4

Source: own work based on enterprise data.

increase in population's remuneration is observed which converges, in turn, with the Christmas period, generally favourable for "investment" shopping in households. The lowest sales volume is recorded, in turn, in second and/or third quarters of the examined years, which may be explained by the fact that in those months money was probably allocated in households for renovations of apartments and holidays, not for purchasing household appliances. It should be remembered, though, that time aggregation resulted in a decreased number of observations in the discussed series from 27 to 9. While identifying seasonal fluctuations, one should base on data from

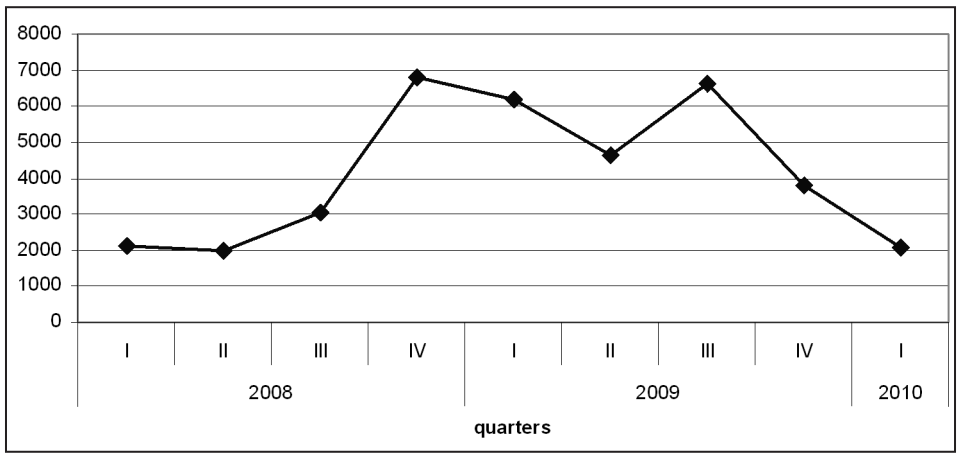

Figure 11. Quarterly sales of colour 5

Source: own work based on enterprise data.

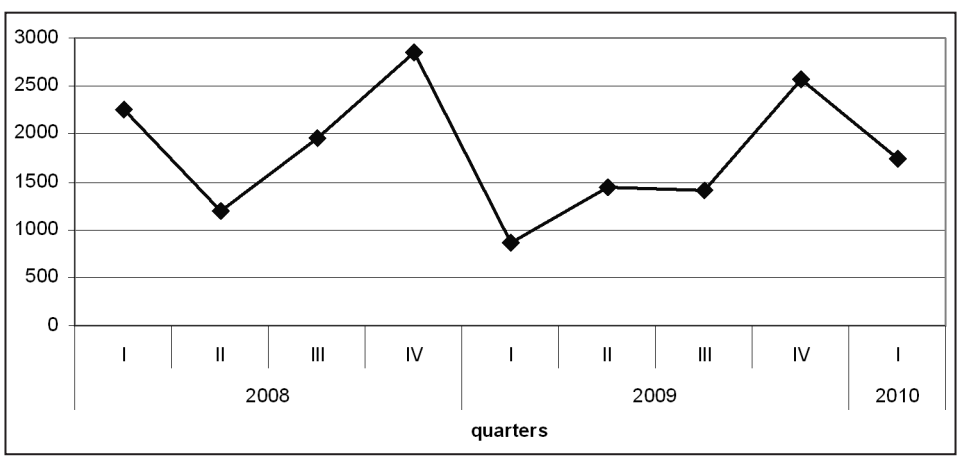

Figure 12. Quarterly sales of colour 6

Source: own work based on enterprise data.

at least three full periods, which would mean three years, i.e. 12 observations in the case of quarterly data. Therefore, in order to fully accept the hypothesis on the occurrence of seasonal fluctuations in the quarterly sales of six product colours, the time series presenting the sales of the product in six colours should be completed by e.g. the remaining data from 2010, which was not possible at the moment of writing this paper. However, the application of the aggregation procedure resulted in the possibility in the case of quarterly sales volume (confirmed by the greater quantity of data) of the product in six colours to employ e.g. the index method or an econometric

model of seasonal fluctuations, and in the case of significantly more data − the Winters model or harmonic analysis.

Since the management of the enterprise was interested also in the overall sales volume of the product (without distinguishing into colours), substantive aggregation was performed, thanks to which the monthly overall product sales volume was obtained from January 2008 to March 2010 (Figure 13). Thus created aggregate showed similar regularities as in the case of the monthly sales volume of the product in six colours.

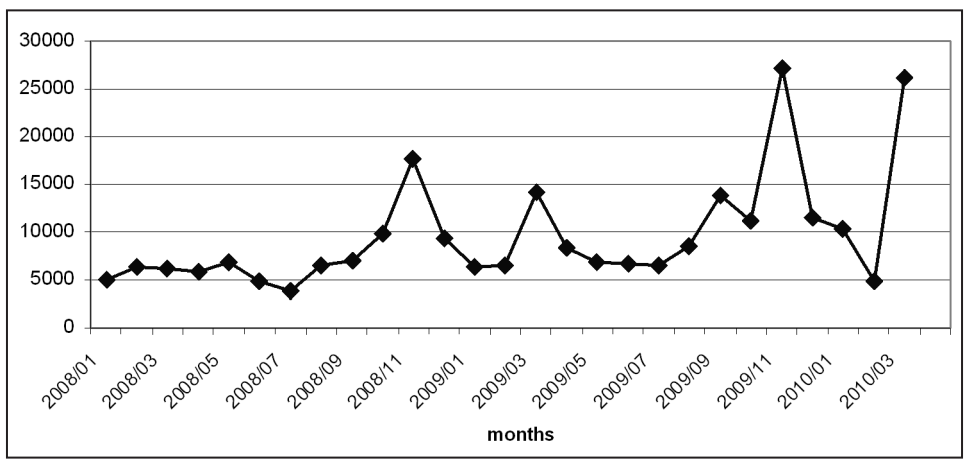

Figure 13. Monthly overall product sales volume

Source: own work based on enterprise data.

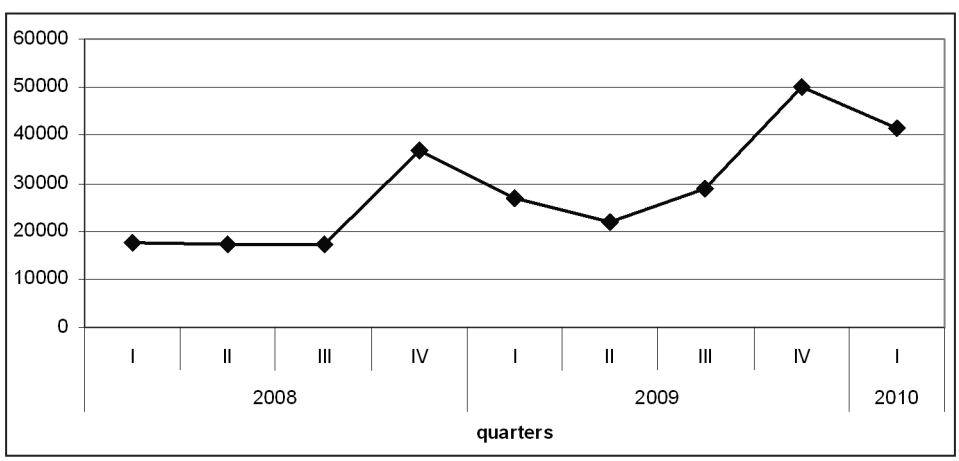

Figure 14. Quarterly overall product sales volume

Source: own work based on enterprise data.

The standard deviation s of the aggregated variable amounted to 5901.9 (units) and the variability coefficient V_s was 61.8%. Therefore, based on this data similar conclusions can be drawn as in the case of data on the sales volume of individual colours of the product. Adjusting the formal model to monthly data may be very difficult, and forecasts established therefrom may be considerably erroneous. Therefore, the decision was taken to look carefully at the quarterly data of the overall product sales volume, which was obtained thanks to the time aggregation of monthly overall sales volume data (Figure 14).

In this case the data shows similar regularities to the data on the quarterly sales volume of individual product colours. Therefore, also in this case such forecasting methods as the index method, seasonal fluctuation econometric model, can be proposed, or – with sufficient amount of data – the Winters model or harmonic analysis, obviously upon confirming seasonal fluctuations statistically in the case of greater numbers of observations.

5. Conclusion

Time aggregation in the presented example enabled obtaining data with explicit regularities from very varied data, and the former data may form a base for constructing the proposed formal forecast models (seasonal fluctuation models) and thereby also the quantitative forecasts of the examined phenomenon. It should be remembered, though, that if deciding to apply such methods, one should have an appropriate number of observations at his disposal, and the established forecasts ought to be short-term.

The "data smoothing" effect is not the case with employing the substantive aggregation – this operation did not enable selecting satisfactory forecast models the application of which would yield good forecasts.

The employment of the proposed seasonal fluctuation models would enable establishing forecasts of quarter's product sales volumes (both in colours and in overall sales). In order to obtain month's forecasts, disaggregation of quarterly forecasts would have to be applied by the top-down procedure, which is the case here. It should be borne in mind while disaggregating that this process should meet the condition of the balancing of forecasts, i.e. the sum of partial forecasts (in the discussed example the sum of monthly forecasts) should be equal to the aggregate (the quarterly forecast). Obviously, disaggregating quarterly forecasts into good monthly forecasts is neither easy nor simple. There are a lot of methods of disaggregating sales forecasts. Among those methods one can distinguish methods using experts' opinions, who are mainly people directly involved in the sales of the disaggregated product. In other methods formal models are used, whose construction is based on the structure of partial variables observed in the past. The following may be distinguished among them [Dittmann 2000, p. 158-171]:

– methods which do not take into account changes in the partial variable structure. In those methods shares of partial variables calculated in different ways are used in the aggregate sales forecast,
– methods which take into account changes in the partial variable structure. In those methods appropriate models regarding the aggregate are constructed (above all trend functions or segment functions), and subsequently depending on the specificity of the examined variable (i.e. taking into account e.g. occurring seasonal fluctuations) and depending on the number of observations describing partial variables and possibilities to use information related to the events attending these sales in the disaggregation process, appropriate modifications and transformations of those models are performed, which as a result enables establishing partial forecasts.

The employment of one of the listed disaggregation methods of quarterly forecasts would require empirical verification – forecasts created as a result of disaggregation should be less erroneous than those created based on original monthly data (otherwise the aggregation procedure and subsequently disaggregation procedure would not make sense whatsoever). Since the issue of forecast disaggregation is quite extensive, the author does not undertake to resolve it in this paper (e.g. due to little data).

In the discussed example the consistency of aggregation is the case since the conclusions drawn based on the aggregated data corresponded to the conclusions resulting from the analysis of non-aggregated (original) data. However, it should be emphasised that this consistency occurs within individual types of aggregation: overall product monthly sales created as a result of substantive aggregation behaved similarly to the monthly sales of individual colours of the product; overall product quarterly sales created as a result of time aggregation behaved similarly to the quarterly sales of individual colours of the product.

Alluding to the statement that by aggregating data part of information on the examined variable is lost in favour of the facilitation or even enabling to solve a problem, it can be noted that in the presented example already after a single (time) aggregation the aim was achieved − finding an appropriate forecasting model which could yield good forecasts. Therefore, it can be presumed that in this case the loss of information was not so significant.

It seems, thus, that based on the conducted tests and conclusions drawn from that a hypothesis may be assumed that data aggregation in certain situations is necessary for being able to construct a good formal forecasting model and as a consequence to establish a good quantitative forecast.

Literature

Aigner D.J., Goldfeld S.M., *Estimation and prediction from aggregate data when aggregates are measured more accurately than their components*, "Econometrica" 1974, Vol. 42, no 1.

Bołt T.W., Krauze K., Kulawczuk T., *Agregacja modeli ekonometrycznych*, PWE, Warszawa 1985.

Dittmann P., *Metody prognozowania sprzedaży w przedsiębiorstwie*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław 2000.
Dittmann P., Szabela-Pasierbińska E., Dittmann I., Szpulak A. *Prognozowanie w zarządzaniu przedsiębiorstwie*, Oficyna Wolters Kluwer, Karaków 2009.
Green H.A.J., *Aggregation in Economic Analysis. An Introductory Survey*, Princeton University Press, New Jersey 1964.
Hanke J.E., Wichern D.W., *Business Forecasting*, Prentice Hall, New Jersey 2005.
Kubiak B., *System Zarządzania Wiedzą we współczesnej organizacji*, Prace i Materiały Wydziału Zarządzania UG, nr 1, Gdańsk 2005.
Muryjas P., Miłosz M., *Współczesne technologie informatyczne*, Wydawnictwo MIKOM, Warszawa 2003.
Słownik 100 tysięcy potrzebnych słów, ed. J. Bralczyk, PWN, Warszawa 2005.
Theil H., *Zasady ekonometrii*, PWN, Warszawa 1979.

ROLA AGREGACJI DANYCH W PROCESIE WYBORU METODY PROGNOZOWANIA SPRZEDAŻY

Streszczenie: W artykule przedstawiono rolę agregacji w procesie prognozowania sprzedaży. Wyjaśniając pojęcie agregacji zauważono, że w prognozowaniu z zagadnieniem tym można mieć do czynienia na etapie analizowania danych prognostycznych oraz na etapie konstruowania prognozy. W pierwszym przypadku agregacja, która chociaż może spowodować utratę części informacji o badanym zjawisku, ułatwia a nawet umożliwia wybór modelu prognostycznego niezbędnego do wyznaczenia prognozy tego zjawiska. W drugim przypadku agregacja umożliwia uzyskanie prognoz dotyczących dłuższych okresów i większych obszarów niż prognozy pierwotnie wyznaczone. Rozważanie teoretyczne zilustrowano przekładem empirycznym, dotyczącym prognozowania sprzedaży artykułów AGD.