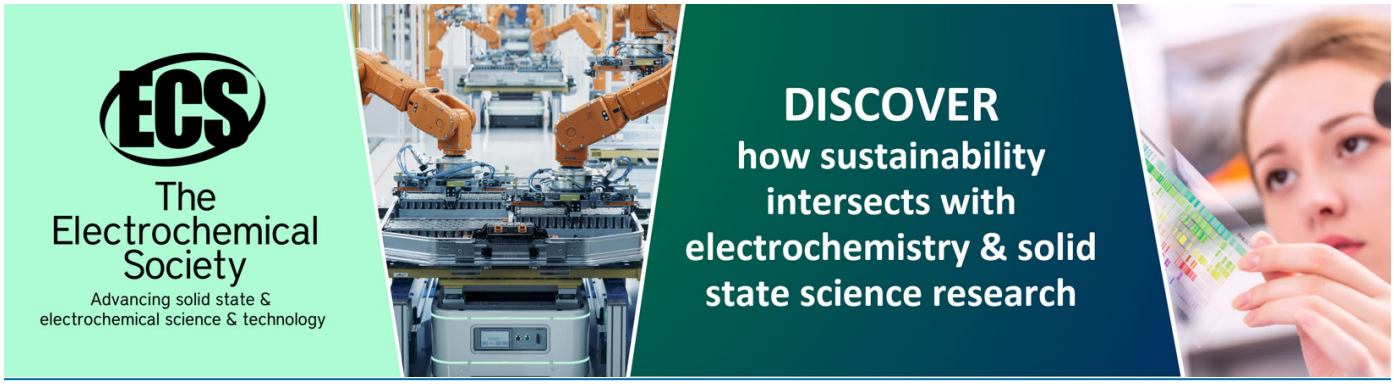
Piecewise Polynomial Aggregation as Preprocessing for Data Numerical Modeling



IOP Conf. Series: Journal of Physics: Conf. Series **1015** (2018) 032028 doi:10.1088/1742-6596/1015/3/032028

Piecewise Polynomial Aggregation as Preprocessing for Data Numerical Modeling

B S Dobronets, O A Popova

Siberian Federal University, Kireskogo 26, Krasnoyarsk, 660074, Russia

E-mail: BDobronets@yandexl.ru

Abstract. Data aggregation issues for numerical modeling are reviewed in the present study. The authors discuss data aggregation procedures as preprocessing for subsequent numerical modeling. To calculate the data aggregation, the authors propose using numerical probabilistic analysis (NPA). An important feature of this study is how the authors represent the aggregated data. The study shows that the offered approach to data aggregation can be interpreted as the frequency distribution of a variable. To study its properties, the density function is used. For this purpose, the authors propose using the piecewise polynomial models. A suitable example of such approach is the spline. The authors show that their approach to data aggregation allows reducing the level of data uncertainty and significantly increasing the efficiency of numerical calculations. To demonstrate the degree of the correspondence of the proposed methods to reality, the authors developed a theoretical framework and considered numerical examples devoted to time series aggregation.

1. Introduction

Aggregation is quite a popular method of converting big data [2, 8, 12]. For example, the application of the histogram allows reducing dimensions of the data set and the level of uncertainty and increasing significantly the efficiency of numerical calculations. It is important to note that the histograms are examples of the symbolic data using in the Symbolic data analysis [2, 3].

Symbolic Data Analysis and Data Mining use the histograms to study a variety of different processes and are applied to model the variability of quantitative characteristics.

Histogram data models and histogram regression models based on the Symbolic analysis are a new important direction to discover knowledge in a data base. Billard L., Diday E. proposed the symbolic data type named as histogram-valued variables to employ them for regression modeling [3, 4].

In this study, the authors propose a new approach to numerical modeling using input data aggregation. To develop a novel method for performing efficient aggregation, let us employ mathematical aggregation functions presented through piecewise polynomial models, including piecewise linear functions and splines. A spline is a good example of the piecewise polynomial functions to perfectly employ in this study. This approach will allow employing the probability density function models as input data.

To examine the structure of input data uncertainty, let us use density functions (DF). It is important that data uncertainty should be studied to identify the relationship between the input and output characteristics when input probability density functions are unknown. The following statements confirm the usefulness of piecewise polynomial models.

The application of the piecewise polynomial functions as a mathematical model allows reducing the level of uncertainty and increasing significantly the efficiency of numerical calculations. This approach allows one to represent accurately enough the arbitrary distribution. It is important to note that despite its simplicity, the piecewise polynomial functions cover all possible ranges of probability density estimation. Histograms and frequency polygons are widely used in practice and are most popular. A histogram is a piecewise constant function which approximates the probability density with accuracy *O(h)*. However, even midpoints of histograms approximate the probability density function with accuracy $O(h^2)$. Consequently, the frequency polygon approximates the function with accuracy $O(h^2)$ [10].

2. Data aggregation

In this section, let us consider the data aggregation as a pre-treatment method for subsequent and numerical modeling. The essence of the aggregation procedure is methods for reducing the dimension of the original empirical data in order to increase the efficiency of data processing, knowledge discovery in data bases and to reduce data uncertainty. Data aggregation plays a most concerned role to extract useful information from large volume of data. The essences of the aggregation procedures constitute methods to reduce the initial data set to less data. Aggregation can be considered as converting data with a high degree of detail to a more generalized representation.

The aggregation procedure has its own advantages and disadvantages. The positive moment is that detailed data are often very volatile due to the impact of different random factors, making difficulties to discover general trends and data patterns. In many cases, it is useful to consider numerical big data in an aggregated form such as summation or average.

It is important to bear in mind that the use of such aggregation procedures as averaging, the exclusion of extreme values (emission), the smoothing procedure can lead to loss of important information. There are various methods of data aggregation. Therefore, the choice of the aggregation method is a complex problem, because the wrong numerical methods of calculation may introduce additional uncertainty, which is not present in the original problem. The data aggregation can used various mathematical models. Numerical probabilistic analysis outlines the following models such as histograms, frequency polygons and splines.

A spline is a sufficiently smooth polynomial function that is piecewise-defined, and possesses a high degree of smoothness at the places where the polynomial pieces connect (which are known as knots). Let us consider the probability density of the random variables as an approximated spline.

In numerical analysis, a Hermite spline is a spline where each piece is a polynomial specified in Hermite form. A cubic Hermite spline is a piecewise polynomial function where each piece is a third degree polynomial specified in Hermite form. On the unit interval of $x \in [0.1]$, polynomial s can be defined by:

$$s(x) = \psi_1(x)s(0) + \psi_2(x)s'(0) + \psi_3(x)s(1) + \psi_4(x)s'(1),$$

where

$$\psi_1(x) = 2x^3 - 3x^2 + 1, \psi_2(x) = x^3 - 2x^2 + x, \psi_3(x) = -2x^3 + 3x^2, \psi_4(x) = x^3 - x^2.$$

In arbitrary interval $[x_i, x_{i+1}]$, the polynomial can be written in the form:

$$s(x) = \psi_1(t)s(x_i) + h\psi_2(t)s'(x_i) + \psi_3(t)s(x_{i+1}) + h\psi_4(t)s'(x_{i+1}),$$

where

$$t = (x - x_i)/h, h = x_{i+1} - x_i.$$

A quintic Hermite spline on uniform mesh ω with step h can be defined by:

$$s(x) = \sum_{i=0}^{n} \phi_0((x - x_i)/h) f(x_i) + \psi_1((x - x_i)/h) f'(x_i) + \psi_2((x - x_i)/h) f''(x_i),$$

where $\phi_i \in C^2$ are basis functions, and if $|x| \geq 1$, then $\phi_i(x) \equiv 0$. If $|x| < 1$, then:

$$\phi_0(x) = (1-|x|)^3(6x^2+3|x|+1), \phi_1(x) = (1-|x|)^3|x|(3|x|+1), \phi_2(x) = (1-|x|)^3x^2.$$

3. Spline aggregation

Let us consider the spline approach to data aggregation. This approach is useful for the following reasons. Underlying this approach is the notion of the spline. The spline can be regarded as a mathematical object that is easy to describe and calculate the mathematical procedures and operations, while maintaining the essence of the frequency distribution of the data. Since the spline is a piecewise polynomial function, then it can be regarded as a data aggregation function in the aggregation issues. An aggregation function performs the numerical calculations on a data set and returns the spline values. Splines are useful for data uncertainty analysis due to the fact that they adequately represent the random distribution of random variables. Despite its simplicity, the spline covers all possible ranges of probability density function estimation. A simple and flexible spline structure greatly simplifies their use in numerical calculations and it has a clear visual image, which is useful for analytical conclusions. It is important to note that the construction of regression models with aggregated inputs requires the use of appropriate numerical procedures. To this end, let us consider numerical probabilistic analysis. Let us use the numerical probabilistic analysis to compute the arithmetic operations for aggregated data and to apply for regression modeling. Let us assume that samples $\Xi = \{\xi_1, \xi_2, ..., \xi_N\}$ of random variable ξ with probability density function *f(x)* and support [*a,b*] are known.

It is remarkable that the histogram stood as the only nonparametric density estimator until the 1950s, when substantial and simultaneous progress was made in density estimation and in spectral density estimation. During the following decade, several general algorithms and alternative theoretical modes of analysis were introduced by Rosenblatt, Parzen, and Cencov [10].

Next, let us consider the use of Richardson's extrapolation to improve the accuracy of the kernel estimator [5].

The basic kernel estimator may be written compactly as [15]:

$$\hat{f}_h(x) = \frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x-\xi_i}{h}) = \frac{1}{Nh}\sum_{i=1}^{N} K_h(x-\xi_i),$$

where $K_h(t) = K(t/h)/h$.

Let us note that:

$$K_h(x,\xi_i) = K(\frac{x-\xi_i}{h}).$$

Here ξ is a random variable with probability density function $f(x)$.

Then the expected value is:

$$\mathrm{E}[\hat{f}(x)] = \mathrm{E}[K_h(x,\xi)]$$

and variability is:

$$\sigma_N = \mathrm{Var}[\hat{f}(x)] = \frac{1}{N}\mathrm{Var}[K_h(x,\xi)].$$

Let us suppose that kernel K satisfies the requirements:

$$\int_{-\infty}^{\infty} K(\eta)d\eta = 1, \int_{-\infty}^{\infty} \eta K(\eta)d\eta = 0$$

and

$$\int_{-\infty}^{\infty} \eta^3 K(\eta)d\eta = 0.$$

Let us denote that:

$$\int_{-\infty}^{\infty} \eta^2 K(\eta)d\eta = \sigma^2.$$

Let us define $f^h(x)$ as:

$$f^h(x) = \mathrm{E}[\hat{f}_h(x)] = f(x) + \sigma^2 h^2 f''(x)/2 + O(h^4). \quad (1)$$

and $f^{2h}(x)$ as:

$$f^{2h}(x) = \mathrm{E}[\hat{f}_{2h}(x)] = f(x) + 4\sigma^2 h^2 f''(x)/2 + O(h^4). \quad (2)$$

Let us apply the Richardson's extrapolation to $f^h(x)$ and $f^{2h}(x)$ [7]. In the next stage, let us multiply (1) by 1/4 to subtract the result from (2). Excluding $\sigma^2 h^2 f''(x)/2$ from (4) and (5), let us get:

$$f(x) = \frac{4}{3} f^h(x) - \frac{1}{3} f^{2h}(x) + O(h^4).$$

Let us note that the approximation to function $f(x)$ has been constructed:

$$f^h_{cor}(x) = \frac{4}{3} \hat{f}^h(x) - \frac{1}{3} \hat{f}^{2h}(x). \quad (3)$$

with accuracy $O(h^4)$.

a b

Figure 1. Improving the accuracy of the probability density function estimation.

In Figure 1, let us represent a numerical example. The solid line is exact probability density function *f(x)*, a − kernel estimator of the probability density function, b − correction of the kernel estimator function by Richardson's extrapolation.

Thus, setting $z \in \omega$ successively, let us obtain values $f^h_{cor}(x_i) = f(x_i) + O(h^4)$. Further, using the obtained values, it is possible to construct systems of linear algebraic equations for constructing a cubic spline [1, 11].

4. The application of the regression approach to spline-aggregated time series

Let us consider the issue of constructing a numerical modeling on aggregated time series. The time series is suitable for representing many practical situations. It should be noted that in many cases, the time-series is analyzed as the data of large amount. To analyze the relationships between the data time series, the authors used the aggregation procedures. To begin with, the following arguments were noted. Although it is well known that the time series describe appropriately the empirical data for many practical and theoretical situations, there is an argument in the studies that time series do not faithfully present phenomena where a set of realizations of the observed variable has a certain degree of variability.

There are two typical situations when this happens [2]: if a variable is measured through time for each individual of a group and the interest does not lie in the individuals but in the group as a whole. In this case, a time series of the sample mean of the observed variable over time would be a weak representation. When a variable is observed at a given frequency (e.g. minutes), but has to be analyzed at a lower frequency (e.g. days).

These two situations describe a contemporaneous and temporal aggregation, respectively. In each case, a time series of distributions would offer a more informative representation than other forms of aggregated time series. As evidence, let us present the viewpoint of Schweizer stating that "distributions are the numbers of the future" [9]. Thus, instead of simplifying them, it seems better to

propose methods which deal with distributions directly. In order to do this, one has to determine how to represent the observed distributions.

In this study, let us propose representing them using a piecewise-polynomial aggregation function, because it offer a good tradeoff between simplicity and accuracy.

Let us consider the use of an regression approach to spline-aggregated time series. In this issue, let us focus on situations where it is necessary to describe the data variability as a regression model. Let *Y* be the spline aggregation variable.

At the beginning, let us propose studying the following model:

$$Y = a_1\varphi_1(t) + a_2\varphi_2(t) + a_3\varphi_3(t),$$

where a_1, a_2, a_3 are constants.

For example, let us consider the temperature data for the last hundred years in Krasnoyarsk city. For each day, from 01 April to 01 October, the data are aggregated in the form of splines Y_i, $i = 1,2,...,184$.

Let us construct approximation empirical cumulative distribution function F_i for each day by quintic Hermite spline *s* [11]:

$$s = a_0\phi_0(t) + a_1\phi_1(t) + a_2\phi_2(t) + \phi_0((x-b)/h),$$

where $t = (x - x_0)/h$, $x_0 = (a+b)/2$, $h = (b-a)/2$; boundary conditions are:

$$s(a) = 0, s'(a) = 0, s''(a) = 0,\ s(b) = 0, s'(b) = 0, s''(b) = 0.$$

Let us find unknown constants a_0, a_1, a_2 by the method of least squares:

$$\Phi(a_0, a_1, a_2) = \min_{a_0, a_1, a_2} \| F_i - s \|^2 .$$

Let us approximate probability density function Y_i by differentiating spline *s*.

In this case, the regression model can be represented in the form:

$$\hat{Y}_i(A) = A_1\varphi_1(t_i) + A_2\varphi_2(t_i) + A_3\varphi_3(t_i), i = 1,2,...,184,$$

where A_1, A_2, A_3 are the probability density functions, φ_1, φ_2, φ_3 are quadratic functions:
$\varphi_1(t_1) = 1, \varphi_1(t_{92}) = 0, \varphi_1(t_{184}) = 0,\ \varphi_2(t_1) = 0, \varphi_2(t_{92}) = 1, \varphi_2(t_{184}) = 0,$
$\varphi_3(t_1) = 0, \varphi_3(t_{92}) = 0, \varphi_3(t_{184}) = 1.$

Density functions A_1, A_2, A_3 are represented in the form of Hermite splines s_i. Splines are defined by mesh $\{x_1^i, x_2^i, x_3^i\}$. Boundary conditions are $s(x_1^i) = 0$, $s'(x_1^i) = 0$, $s(x_3^i) = 0$. For variables, x_1^i, x_3^i are chosen by regression curves of minimum and maximum temperatures. x_2^i is chosen by the regression curve of average temperatures and:

$$\Phi(A) = \sum_{i=1}^{184} \rho^2(\hat{Y}_i(A), Y_i),$$

$$\Phi(A^*) = \min_A \Phi(A).$$

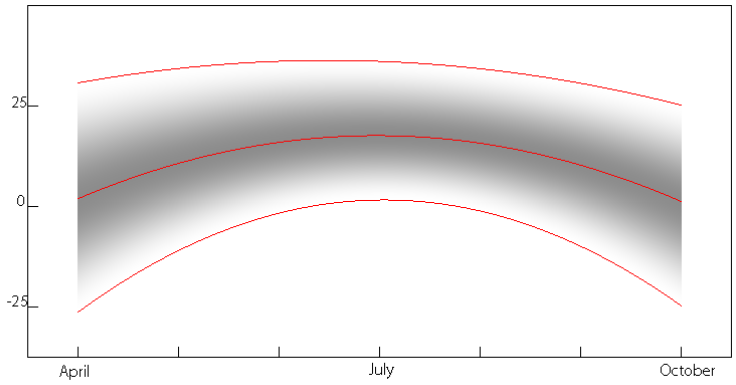


Figure 2. Probability density functions of the temperature data for the last 70 years in Krasnoyarsk city

Figure 2 shows the probability density functions of the temperature data for the last 70 years in Krasnoyarsk city, from 01 April to 01 October. Shades of gray define the values of the probability density. The top and bottom lines represent maximum and minimum temperature on each day over the last 70 years respectively. Middle line denotes the mean of daily temperature over the last 70 years. Each vertical section is approximation of the probability density function of the temperature corresponding to a certain day of the year, according to the observations of the day in the last 70 years. At the first stage, the data presented the cumulative distribution function for each day in a quintic Hermite spline as it was mentioned above.

The regression data are presented in the form of a derivative of a quintic Hermite spline. Thus, the temperature data for last 70 years from April to October are aggregated with the help of quintic Hermitian splines. The visual representation shows the change in the maximum, minimum, and most probable temperature. Shades of gray show the distribution of the probability density.

5. Conclusion

Although there are many ways of data aggregation, including simple average, let us argue that the use of piecewise linear and piecewise polynomial aggregation functions will offer a more informative representation of the variability in the data than other forms of data aggregation. To prove their thesis, the authors considered the aggregation procedure based on the histogram time series. Using these types of data aggregation for preprocessing and regression modeling, it is possible to contribute to the reliability of the study of natural systems and processes. The spatial and time aggregation procedures help to reduce the amount of computation in data processing and are an important basis for the extraction of useful knowledge from large volumes of data. Developed methods reduce the level of uncertainty in the information flow; reduce significantly the processing time and the implementation of numerical procedures. This approach allows one to choose the mode of interactive visual modeling to provide the necessary data for operational decision making under remote surveillance techniques and distributed object systems. In concluding the discussion about the applicability of this approach to practice, it is necessary to mention the advantage of uncertainty treatment and big data processing. Using the proposed model, applications with real and simulated data are presented.

References

[1] Ahlberg J H, Nilson E N, Walsh J L 1967 *The theory of splines and their applications*. (Academic Press, New York)
[2] Arroyoa J, Maté C 2009 Forecasting histogram time series with k-nearest neighbours methods *International Journal of Forecasting* 192–207
[3] Billard L, Diday E 2006 *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. (Wiley)
[4] Dias S, Brito P 2015 *Linear Regression Model with Histogram-Valued Variables*. Retrieved from: wileyonlinelibrary.com DOI:10.1002/sam.11260

[5] Dobronets B S, Popova O A 2017 Improving the accuracy of the probability density function estimation. *Journal of Siberian Federal University — Mathematics and Physics* **1** 16–21
[6] Dobronets B, Popova O 2016 Numerical Probabilistic Approach for Optimization Problems. Scientific Computing, Computer Arithmetic, and Validated Numerics. *Lecture Notes in Computer Science* **9553** 43–53
[7] Marchuk G I, Shaidurov V V 1983 *Difference methods and their extrapolations* (Springer--Verlag, New York)
[8] Nava J 2012 A New Justification for Weighted Average Aggregation in Fuzzy Techniques, *Journal of Uncertain Systems* **6(2)** 84–85
[9] Schweizer B 1984 Distributions are the numbers of the future, *Proceedings of the mathematics of fuzzy systems meeting* (Naples, Italy) pp. 137–149
[10] Scott R W 2015 *Multivariate density estimation: theory, practice, and visualization* (John Wiley & Sons, New York)
[11] Späth H 1995 *One Dimensional Spline Interpolation Algorithms* (Taylor & Francis)
[12] Tchangani A 2013 Bipolar Aggregation Method for Fuzzy Nominal Classification Using Weighted Cardinal Fuzzy Measure (WCFM) *Journal of Uncertain Systems* **7(2)** 138–151