

### **5. Заключение.**

*Автор считает, что в данной работе новыми являются постановка задачи планирования производства в условиях неопределенности качества готовой продукции и алгоритмы решения возникающих при этом вычислительных подзадач.*

Первая задача определения максимального объема заказа, который можно назначить на литейный агрегат, решается с помощью итеративного алгоритма. Конечность числа итераций и сходимости процедуры доказываются. Вторая задача определения мощностей, необходимых для выполнения очередного заказа, решается аналитически.

### **Литература**

1. LaValle S.M. Planning Algorithms // Cambridge University Press. 2006. 842 p.
2. O'Kane J.M., Tovar B., Cheng P., LaValle S.M. Algorithms for Planning Under Uncertainty in Prediction and Sensing // Autonomous Mobile Robots: Sensing, Control, Decision-Making and Applications. Series in Control Engineering, chapter 13, 2006, pp. 501-547.
3. J. Blythe An Overview of Planning Under Uncertainty // AI Magazine, 20(2), 1999, pp. 37-54.
4. Аристова Н.И. Управление уровнем автоматизации при производстве роботов, идеальная и вероятностная схемы // Автоматика и телемеханика, 2014. № 9. С. 122-132.
5. Потапов М.А., Нелюбин А.П., Соловьёв И.С., Павлов А.А. Решение задач управления производством с большим объемом спецификаций // Информационные технологии в науке, образовании и управлении: труды межд. конф. IT + S&E'16 (Гурзуф, 22.05–01.06.2016 г) / под. ред. проф. Е.Л. Глоризова. М.: ИНИТ, 2016. Весенняя сессия. С. 105-109.

### **Solution of manufacture planning problem under uncertainty of the end-product quality**

*Andrey Pavlovich Nelyubin, junior research scientist, Mechanical Engineering Research Institute of the RAS*

*The article presents problem definition of manufacture planning under uncertainty of the end-product quality. The overall approach of solution of this problem is proposed and numerical algorithms of solutions of arising problems are described.*

*Key words: optimization, scheduling theory, manufacture planning, uncertainty.*

УДК: 519.24

## **О ПОДХОДАХ К АГРЕГАЦИИ ДАННЫХ В ЗАДАЧАХ ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТЕЙ**

*Ольга Аркадьевна Попова, канд. техн. наук, доцент,*

*e-mail: olgaarc@yandex.ru,*

*Сибирский федеральный университет,*

*Институт космических и информационных технологий,*

*http://www.sfu-kras.ru*

*Рассмотрены новые подходы преобразования данных для задач восстановления функциональных зависимостей на основе кусочно-полиномиальной агрегации. Предлагается подход к построению регрессионных зависимостей на основе агрегированных временных рядов в виде эрмитовых кубических сплайнов. Исследованы вопросы точности агрегации.*

*Ключевые слова: численный вероятностный анализ, агрегация, кусочно-полиномиальные модели, функции плотности.*

### **Введение**

Проблема восстановления зависимостей по эмпирическим данным уже не первое десятилетие широко обсуждается в научных кругах, поскольку имеет важное прикладное значение [1,5]. В статье рассматриваются вопросы исследования изменчивости эм-

пирических данных больших объёмов при отсутствии знаний о виде восстанавливаемой зависимости. Известным методом решения данной задачи является подход, основанный на применении регрессионных моделей. Предлагается метод построения регрессионных зависимостей на агрегированных данных. Агрегирование данных применяется как метод предварительной обработки эмпирических данных для последующего численного моделирования. Известно, что агрегацию часто используют в задачах анализа данных, когда необходимо перейти от данных с высокой степенью детализации к более обобщённому представлению. Примером таких процедур является простое суммирование, вычисление среднего, медианы, диапазона максимальных или минимальных значений т.е. интервальных данных, построение функции распределения.

Процедура агрегирования имеет свои преимущества и недостатки. С положительной стороны, мы отмечаем, что подробные данные часто являются очень неустойчивыми из-за воздействия различных случайных факторов, затрудняя обнаружение общих тенденций и шаблонов данных. Во многих случаях полезно рассматривать большие числовые данные в агрегированной форме, такой как суммирование или среднее. Важно иметь в виду, что использование таких процедур агрегирования, как усреднение, исключение экстремальных значений (эмиссия), процедура сглаживания может привести к потере важной информации. Поэтому выбор метода агрегирования является важной задачей, поскольку без предварительного исследования легко получить дополнительную неопределённость, которой нет в исходной постановке.

Для агрегирования данных используются различные математические модели. В тех случаях, когда данные могут быть представлены частотными распределениями рассматриваемых характеристик или признаков, предлагается использовать кусочно-полиномиальные модели. Частным примером кусочно-полиномиальных моделей является гистограмма, которая представляет собой кусочно-постоянную функцию агрегирования. Гистограмма, с точки зрения процесса агрегирования, во многих случаях представляет собой альтернативу операциям усреднения или построения интервальных данных. В отличие от указанных операций применение гистограмм позволяет повысить точность вычисления за счёт использования информации о частотном распределении данных вместо замены набора данных одним значением, например, значением выборочного среднего или моды. Применение гистограммы позволяет уменьшить размерность набора данных, снизить уровень неопределённости и значительно повысить эффективность численных расчётов. Важно отметить, что гистограммы являются примерами использования символьных данных, понятие которых рассматривается в Символьном анализе данных, наравне с интервалами. Billard, L., Diday, E. предложили символьный тип данных, называемый гистограммными переменными, для использования их в регрессионном моделировании [5]. Понятие гистограммно-значимой переменной используется для построения гистограммных регрессионных моделей, что является новым важным направлением для теоретических исследований и решения практических задач обнаружения зависимостей в базе данных. Кроме гистограмм для агрегации полезно рассмотреть полиграммы, частотные полигоны, сплайны. Важно отметить, что, несмотря на свою простоту, кусочно-полиномиальные функции охватывают всевозможные диапазоны оценки функции плотности вероятности. Важно также отметить некоторые их свойства. Например, гистограмм, представляя собой кусочно-постоянную функцию, аппроксимирует плотность вероятности с точностью  $O(h)$ . Однако уже средние точки гистограмм аппроксимируют функцию плотности вероятности с более высокой точностью  $O(h^2)$ . Следовательно, частотный полигон аппроксимирует функцию с точностью  $O(h^2)$ .

#### **Сплайн-подход к агрегированию данных**

Помимо гистограмм и частотных полигонов в качестве математических моделей агрегатов можно использовать сплайны, как кусочно-полиномиальные функции.

Сплайн представляет собой достаточно гладкую кусочно - полиномиальную функцию.

Рассмотрим сплайн-подход к агрегированию данных. Этот подход полезен по следующим причинам. Поскольку сплайн является кусочно-полиномиальной функцией, то его можно рассматривать как функцию агрегирования данных. Функция агрегирования выполняет численную обработку наборов данных и возвращает сплайн-значения. Сплайны полезны для анализа неопределённости в данных из-за того, что они адекватно представляют частотное распределение данных.

Предположим, что нам известна выборка  $\Xi = \{\xi_1, \xi_2, \dots, \xi_N\}$  случайной величины  $\xi$  с функцией плотности вероятности  $f(x)$  и носителем  $[a, b]$ . В [2] рассматривается новый подход к восстановлению функция плотности вероятности на основе эмпирических данных. Подход основан на аппроксимации функции плотности вероятности в некоторой точке с использованием прямоугольных ядер переменной ширины  $h$ . Важно отметить, что подход имеет определённое сходство с гистограммами, в частности, с методом усреднения гистограмм. С другой стороны, структура и точность построенных оценок соответствуют ядерному методу. Таким образом, используя данные  $\Xi$  следуя работе [2] можно получить оценку  $\hat{f}(z)$  функции плотности  $f(z)$ , такую что

$$|f(z) - \hat{f}(z)| \approx h^4, \forall z \in [a, b],$$

где  $h$  — параметр сглаживания.

Рассмотрим вопрос построения сплайна  $s$ , аппроксимирующего функцию плотности  $f(x)$ , так чтобы выполнялась оценка

$$\|f - s\| \leq Ch^4.$$

Для этих целей построим в области  $[a, b]$  сетки  $\omega_z = \{z_i = a + ih_z, i = 0, \dots, N_z\}$ ,  $\omega_x = \{x_i = a + ih_x, i = 0, \dots, N_x\}$ . На сетке  $\omega_z$  вычислим значения  $f_i = \hat{f}(z_i)$ . Сплайн  $s$  будем строить на сетке  $\omega_x$ . Краевые условия выберем следующим образом  $s(a) = 0, s'(a) = 0, s(b) = 0, s'(b) = 0$

$$\sum_{i=1}^{N_z} (s(z_i) - f_i)^2 \rightarrow \min. \tag{1}$$

В случае кубических сплайнов, как классических так, и эрмитовых, задача (1) сводится к решению пятидиагональной системы линейных алгебраических уравнений.

Для кубических сплайнов справедлива следующая оценка

$$\|f^\nu - s^\nu\| \leq Kh_x^{4-\nu} \|f^{(4)}\|, \tag{2}$$

где  $K$  — константа, не зависящая от  $h_x$ . Задачу можно упростить, если вычислить в узлах сетки  $\omega_x$  значения  $\hat{f}_i$ .

Для этих целей будем использовать значения  $\hat{f}_{cor}(z_i)$  и процедуры сглаживания. Например, для классических кубических сплайнов можно использовать метод скользящего среднего, метод взвешенной локальной регрессии, фильтр Савицкого–Голея. Следует стремиться, чтобы выполнялись оценки

$$|\hat{f}_i - f(x_i)| = O(h_x^4).$$

В этом случае построение сплайна сводится к решению трех диагональной системы линейных алгебраических уравнений и будет выполнена оценка (2).

Для эрмитовых кубических сплайнов в узлах сетки  $\omega_x$  необходимо вычислить  $\hat{f}_i$  и значения производных  $\hat{f}'_i$ . Будем использовать фильтр Савицкого–Голея с кубическими полиномами. В этом случае для построения эрмитовых кубических сплайнов до-

статочны локальных вычислений. На каждом интервале  $[x_{j-1}, x_j], j = 1, \dots, n$  эти сплайны представимы в виде [3]

$$s(x) = f_{j-1}v((x-x_{j-1})/h_j) + f'_{j-1}w((x-x_{j-1})/h_j) + \\ + f_jv((x-x_j)/h_j) + f'_jw((x-x_j)/h_j),$$

где  $v(x) = (|x|-1)^2(2|x|+1); w(x) = x(|x|-1)^2$ .

Важно отметить, что построение регрессионных моделей с агрегированными входными данными требует использования соответствующих числовых процедур. С этой целью применяется численный вероятностный анализ (ЧВА). Отличительной особенностью ЧВА является наличие развитых арифметических операций над функциями плотности вероятности, для которых вводится понятие ФПВ – значные переменные. В рамках ЧВА имеется возможность вычисления функций от случайных аргументов с использованием процедур построения вероятностных расширений. В рамках ЧВА решаются различные задачи численного анализа, в том числе задачи интерполяции, аппроксимации и оптимизации.

**Построение регрессионных моделей для агрегированных временных рядов**

Рассмотрим пример построения числовой модели для агрегированных временных рядов. Временной ряд хорошо подходит для представления многих практических ситуаций. Следует отметить, что во многих случаях временные ряды анализируются как большие данные. Для анализа взаимосвязи между временными рядами данных мы используем процедуры агрегирования.

Известно, что временные ряды хорошо описывают эмпирические данные для многих практических и теоретических ситуациях, тем не менее, существуют исследования, в которых утверждается, что временные ряды неверно представляют явления, в которых набор реализаций наблюдаемой переменной имеет определенную степень изменчивости. Существуют две типичные ситуации, когда это происходит. Первая ситуация имеет место, если переменная измеряется во времени для каждого индивида группы, и интерес исследователя относится не к индивидам в отдельности, а к группе в целом. В этом случае временной ряд выборочного среднего наблюдаемой переменной во времени будет слабым представлением. Вторая ситуация, когда переменная наблюдается на данной частоте (например, минуты), но ее необходимо анализировать с меньшей частотой (например, дней). Эти две ситуации описывают распределенное и временное агрегирование, соответственно. В каждом случае временные ряды распределений предлагают более информативное представление, чем другие формы агрегированных временных рядов [4].

Для описания восстанавливаемой зависимости временного ряда рассмотрим модель, представляющую собой линейную комбинацию произведений неизвестных параметров и базисных функций.

Рассмотрим данные о температуре за последние сто лет в городе Красноярске. Для каждого дня с 01 апреля по 01 октября данные агрегируются в виде эрмитовых кубических сплайнов. В этом случае модель регрессии может быть представлена в виде

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

где  $A_1, A_2, A_3$  — представляют собой неизвестные функции плотности вероятности, а функции  $\varphi_1, \varphi_2, \varphi_3$  представлены в следующем виде:

$$\varphi_1(t_1) = 1, \varphi_1(t_{92}) = 0, \varphi_1(t_{184}) = 0, \\ \varphi_2(t_1) = 0, \varphi_2(t_{92}) = 1, \varphi_2(t_{184}) = 0, \\ \varphi_3(t_1) = 0, \varphi_3(t_{92}) = 0, \varphi_3(t_{184}) = 1.$$

В качестве моделей, используем эрмитовы кубические сплайны. Сплайн определяется на сетке  $\{x_1^i, x_2^i, x_3^i\}$ . Граничные условия имеют вид  $s(x_1^i) = 0, s'(x_1^i) = 0, s(x_3^i) = 0, s'(x_3^i) = 0$ . Кроме того  $s'(x_2^i) = 0$  и значение  $s(x_2^i)$  выберем из условия

$$\int_{x_1^i}^{x_3^i} s(\xi) d\xi = 1.$$

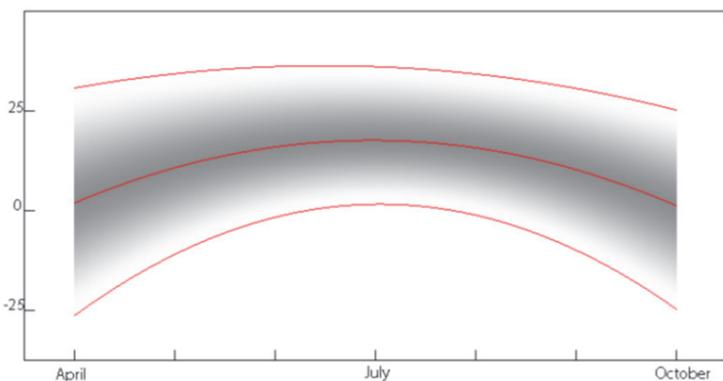
Для нахождения  $A_1, A_2, A_3$  потребуем выполнения условия оптимальности

$$\sum_{i=1}^{184} \rho^2(Y_i, \hat{Y}_i) \rightarrow \min.$$

На рис. 1 представлены функции плотности вероятности данных о температуре за последние сто лет в городе Красноярске с 01 апреля по 01 октября. Оттенки серого определяют значения функции плотности вероятности. Верхняя и нижняя линии показывают максимальную и минимальную температуру в день за последние 100 лет соответственно. Средняя линия обозначает среднее значение суточной температуры за последние сто лет. Каждое вертикальное сечение представляет собой аппроксимацию функции плотности вероятности температуры, соответствующей определенному дню года, согласно наблюдениям дня за последние сто лет. На первом этапе агрегирования данные представлены для каждого дня в виде сплайна. Данные регрессии представлены в виде кубических сплайнов Эрмита. Таким образом, данные о температуре за сто лет с апреля по октябрь агрегируются с помощью кубических эрмитовых сплайнов. Визуальное представление показывает изменение максимальной, минимальной и наиболее вероятной температуры. Оттенки серого показывают распределение плотности вероятности.

**Заключение**

В то время, как существует много способов агрегирования данных, включая простое среднее, использование кусочно-полиномиальных функций агрегации предлагает более информативное представление существующих зависимостей в данных, чем другие формы агрегирования.



**Рис.1. Агрегация температурных значений в г. Красноярске за последние сто лет**

В работе рассматриваются вопросы применения сплайн-агрегации как более информативные модели для представления и анализа изменчивости данных в задачах восстановления зависимостей, чем другие формы агрегирования. С этой целью мы рассмотрели процедуру агрегирования, основанную на представлении временных рядов, используя сплайн-агрегацию данных.

Такие процедуры агрегирования помогают сократить объем вычислений при обработке данных и являются важной основой для извлечения полезных знаний из больших объемов данных. Обладая более высоким порядком сходимости, чем гистограммы и кусочно-линейные функции, сплайны повышают точность вычислений. Разработанные методы снижают уровень информационной неопределенности и существенно сокращают время обработки данных и выполнение численных процедур.

Автор считает, что новыми результатами являются применение кусочно-полиномиальных моделей к агрегации данных и метод построения регрессионных зависимостей на основе численного вероятностного анализа.

**Литература**

1. Добронетц Б.С., Попова О.А. Численный вероятностный анализ неопределённых данных Сибирский федеральный университет, Институт космический и информационных технологий. Красноярск, 2014.
2. Dobronets B.S., Popova O.A. Improving the accuracy of the probability density function estimation (2017) Journal of Siberian Federal University — Mathematics and Physics, 10 (1). pp. 16-21.
3. Dobronets B., Popova O. Numerical probabilistic approach for data nonparametric analysis // Applied methods of statistical analysis. Nonparametric approach. Proceedings of the international workshop. 2015. С. 376-384.
4. Dobronets B.S., Popova O.A. The numerical probabilistic approach to the processing and presentation of remote monitoring data // Журнал Сибирского федерального университета. Серия: Техника и технологии. 2016. Т. 9. № 7. С. 960-971.
5. Billard L., Diday E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley.

**About approaches to the data aggregation in dependence restoration problems**

*Olga Arkadevna Popova, Ph.D., Associate Professor Computer Science Dept., Institute of Space and Information Technology, Siberian Federal University*

*New approaches to data processing for the functional dependence recovery on the basis of piecewise polynomial aggregation are considered. An approach is proposed for constructing regression dependencies on the basis of aggregated time series in the form of Hermitian cubic splines. Questions of the accuracy of aggregation are studied.*

*Keywords: numerical probability analysis, piecewise polynomial models, spline aggregation, functional dependence recovery, regression modeling.*

УДК 519.816

**СИСТЕМА ДЛЯ ВЫБОРА И АНАЛИЗА МНОГОКРИТЕРИАЛЬНЫХ РЕШЕНИЙ С НЕПОЛНЫМИ ПРЕДПОЧТЕНИЯМИ**

**Владислав Владимирович Подиновский**, д-р техн. наук, профессор,  
*e-mail: podinovski@mail.ru,*

*Национальный исследовательский университет «Высшая школа экономики»,  
[http://www.hse.ru,](http://www.hse.ru)*

**Михаил Андреевич Потапов**, канд. физ.-мат. наук, вед. научн. сотр.,  
*e-mail: pma@com2com.ru,*

*Институт автоматизации проектирования РАН,  
[http://www.icad.org.ru,](http://www.icad.org.ru)*

**Андрей Павлович Нелюбин**, мл. научн. сотр.,  
*e-mail: nelubin@gmail.com,*

*Институт машиноведения РАН,  
[http://www.imash.ru,](http://www.imash.ru)*

**Иван Сергеевич Соловьёв**, мл. научн. сотр.,  
*e-mail: ivan.solovyev@phystech.edu,*

*Институт автоматизации проектирования РАН,  
<http://www.icad.org.ru>*

**Александр Александрович Павлов**, мл. научн. сотр.,  
*e-mail: ckidoz@gmail.com,*

*Институт автоматизации проектирования РАН  
<http://www.icad.org.ru>*

*В докладе представлены варианты реализации согласительных решений для многокритериальных задач выбора, которые анализируются методами теории важности критериев и*