УДК 004.94:658 DOI: 10.31673/2412-4338-2018-0-2-46-59

Курченко О. А., **Терещенко А. И.** *Государственный университет телекоммуникаций, Киев*

КЛАСТЕРИЗАЦИЯ ДАННЫХ МНОГОМЕРНЫХ КРИТИЧЕСКИХ ПАРАМЕТРОВ ПРОЦЕССА ПРОИЗВОДСТВА ДЛЯ ОЦЕНКИ ФАКТОРНОГО ВЛИЯНИЯ НА АТРИБУТЫ КРИТИЧЕСКОГО КАЧЕСТВА ПРОДУКТА

Предложен метод многомерного статистического анализа для оценивания характера и влияния многомерных массивов временных данных критических параметров процесса производства на атрибуты критического качества продукта. Рассмотрен объектно-ориентированный подход к кластеризации исследуемых массивов данных для проведения факторного анализа. Метод рекомендовано применять на этапе разработки продукта и проектирования производства.

Ключевые слова: *функция качества, менеджмент качества, кластеризация, многомерные данные, стратегия QbD, процессный подход, структурирование функции качества (QFD).*

Kurchenko O. A., Tereshchenko A. I. *State University of Telecommunications, Kyiv*

DATA CLUSTERING OF MANUFACTURE PROCESS MULTIDIMENSIONAL CRITICAL PARAMETERS FOR EVALUATING THE FACTOR'S INFLUENCE ON PRODUCT CRITICAL QUALITY ATTRIBUTES

Nowadays, competitiveness and efficiency of companies must be continuously improved to face worldwide competitors therefore the product design process require new methodologies of computer-aided development. This article presents an approach for a model-based planning process for the early phases of manufacturing system planning (MSP). The main goals are a better integration of MSP with product development in the early design phases. The presented approach is based on object-oriented modeling and is supported by a modeling scheme which uses the object-oriented modeling language (R). The article suggests a method of multivariate statistical analysis of the affect of critical process parameters (CPPs) and its factors, on product critical quality attributes (CQAs) with data clustering. It is proposed to transform clusters of critical process parameters into influence factors, which increases the number of degrees of freedom for multivariate statistical research. Method propose the clustering of contingency data array based on multivariate statistical analysis (MSA) for evaluating the influence of critical process parameters factors on the time multivariate product critical quality attributes. Factorized multivariate CPPs increase the possibility to use methods of multivariate statistical analysis for evaluating the influence of CPPs on the multivariate CQAs. The proposed method presents an actual information technology for assessing the affect of time multidimensional data objects and separate components of critical process parameters on product critical quality attributes.

Keywords*: quality function, quality management system, data clustering, multidimensional data, QbD, Quality-by-Design strategy, process approach, quality function deployment (QFD), statistical analysis.*

Курченко О. А., Терещенко А. І. *(Державний університет телекомунікацій, Київ)*

КЛАСТЕРИЗАЦІЯ ДАНИХ БАГАТОВИМІРНИХ КРИТИЧНИХ ПАРАМЕТРІВ ПРОЦЕСУ ВИРОБНИЦТВА ДЛЯ ОЦІНКИ ФАКТОРНОГО ВПЛИВУ НА АТРИБУТИ КРИТИЧНОЇ ЯКОСТІ ПРОДУКТУ

У наш час конкурентоспроможність і ефективність компаній повинні постійно вдосконалюватися, щоб протистояти у світовій конкуренції, тому процес розробки продукту вимагає нових методологій комп'ютерної підтримки проектування і інтенсивної інтеграції чисельного моделювання. У статті представлений підхід до процесу планування на основі моделей для ранніх етапів планування виробничої системи (MSP, manufacturing system planning). Однією з основних цілей є поліпшення інтеграції MSP з розробкою продукту (PD, product development) на ранніх етапах розробки і підвищення координації об'єктно-орієнтованих дисциплін планування, пов'язаних з MSP.

Представлений підхід заснований на об'єктно-орієнтованому моделюванні і підтримується схемою моделювання, в якій використовується мова об'єктно-орієнтованого моделювання (R). Пропонується метод багатовимірного статистичного аналізу впливу критичних параметрів процесу (CPPs, critical process parameters) і їх факторів на атрибути критичної якості продукту (CQAs, critical quality attributes) з кластеризацією даних. Пропонується перетворити кластери критичних параметрів процесу у фактори впливу, що збільшує кількість ступенів свободи для багатовимірних статистичних досліджень. Метод пропонує використовувати кластеризацію масивів випадкових даних на основі багатовимірного статистичного аналізу (MSA, multivariate statistical analysis) для оцінки впливу факторів критичних параметрів процесу на багатовимірні атрибути критичної якості продукту. Факторизовані багатовимірні CPPs збільшують можливість використання методів багатовимірного статистичного аналізу для оцінки впливу CPPs на багатовимірні CQAs. Цей метод являє собою фактичну інформаційну технологію для оцінки впливу часових багатовимірних об'єктів даних і окремих компонентів критичних параметрів процесу на атрибути критичної якості продукту.

Ключові слова: *функція якості, управління якістю, кластеризації даних, багатовимірні дані, стратегія якості за методом розробки, технологічний підхід, функція якості, статистичний аналіз.*

1. Введение. В настоящее время процессный подход является доминирующей концепцией в управлении и контроле взаимосвязей между процессами, а также взаимодействий между функциональными уровнями организации при достижении стандартизированного качества продукции [1], которую можно трактовать как ценность для потребителя. Главное преимущество процессного подхода в системе менеджмента качества QMS (Quality Management System) состоит в синхронизированном управлении и контроле цепочкой производственных процессов (горизонтальная линия управления) и отношений между функциональными уровнями организации (вертикальная линия управления).

В статье внимание уделено важной для приложений информационных технологий горизонтальной системе сети процессов, где на технологическом уровне производства осуществляется управление параметрами процесса для выполнения требований к характеристикам продукции [2, 3]. При этом важнейшей задачей для QMS является определение характера изменчивости критических атрибутов качества продукта *CQAs* (Critical Quality Attributes) от изменчивости и влияния критических параметров процесса производства *CPPs* (Critical Process Parameters) по всей горизонтальной цепи для управления созданием ценности/качества для потребителя [4, 5].

Известен метод Г. Тагути (Genichi Taguchi) решения этой задачи путём дробного предикторного планирования эксперимента с субъективными тестовыми наборами параметров проектирования и факторов помех [6]. Эти данные формируют матрицу ортогональных расположений. По результатам воздействия факторов помех на группы ортогональных параметров проектирования оценивают их робастность и анализируют возможные потери квадратичной функции качества.

В отличие от такого решения предложено применить компьютерно-интенсивные информационные технологии многомерного статистического анализа (MSA, Multivariate Statistical Analysis) для проведения статистического полнофакторного эксперимента с непосредственной классификацией каждой из таблиц наблюдений после проведения разведочного анализа данных. При этом статистические оценки многомерного факторного анализа адекватно дополняются оценками компонентного влияния *CPPs* на *CQAs*.

Для применения информационных технологий при решении каузальной задачи $CPPs \rightarrow CQAs$ для QMS организуется система сбора данных о ходе выполнения процесса в виде критических атрибутов качества продукта (*CQAs*) и критических параметров процесса (*CPPs*) [7]. Существенно, что зависимость *CQAs* от *CPPs* рассматривается как функция качества: $CQAs = F(CPPs)$.

Актуальность предлагаемого метода состоит в использовании информационных технологий многомерного статистического анализа для оценки факторного влияния временных многомерных критических параметров процесса производства (*CPPs*) на временные критические атрибуты качества продукта (*CQAs*). Сформированные по этим

оценкам требования к технологическим параметрам процесса производства учитываются при мониторинге и при составлении стандартных рабочих инструкций для каждого шага производственного процесса.

2. Цели и задачи. Для поэтапного преобразования требований потребителя в критические атрибуты *CQAs* и *CPPs*, а также способы контроля (мониторинга) технологических операций производства функция качества структурируется по принципу QFD (Quality Function Deployment) [8]. В результате устанавливают индикаторы (критерии, атрибуты) выполнения важнейших (критических) процессов и формализуют требования к функции качества.

Аналитической методологией выполнения формализованных требований к заданной функции качества процесса на уровне применения информационных технологий является набор приёмов и способов «конструирования качества» QbD (Quality-by-Design) [4, 5]. QbD использует аналитические технологии PAT (Process Analytical Technology) [9, 10] для анализа связей критических параметров процесса производства и их влияния на критические атрибуты качества продукта. Методы организации такого анализа при QbD аналогичны методам построения матриц взаимосвязи при QFD и создают основу конвергенции подходов и приёмов QFD и QbD при достижении качества продукта, как показано на рис. 1.

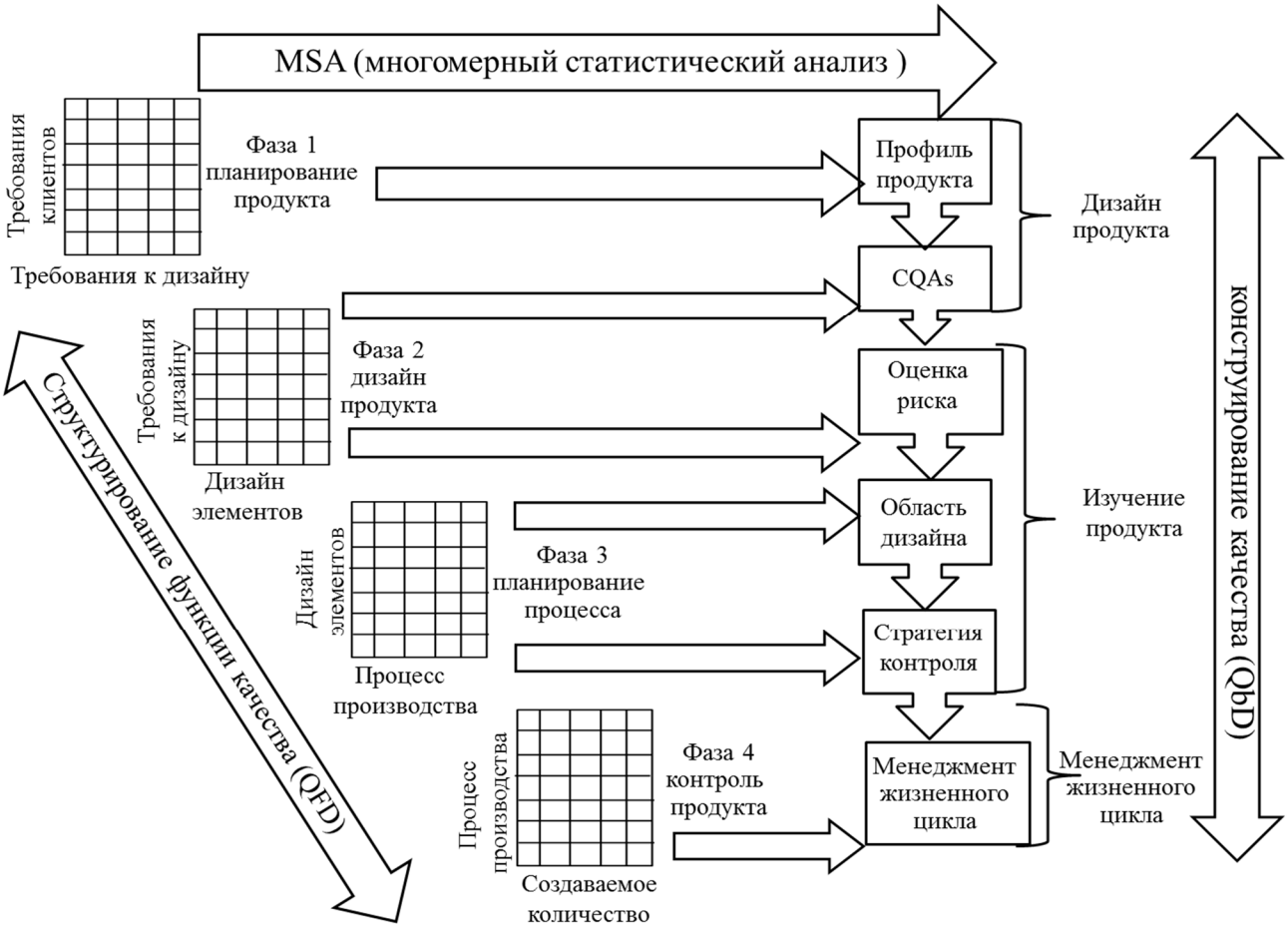


Рис. 1. Структура конвергенции технологий QFD и QbD при достижении качества продукта

Как показано на рис. 1 каждая фаза QFD может рассматриваться как объект применения аналитического аппарата QbD. Существенно, что методология QbD направлена на обеспечение качества продукта на ранних этапах проектирования, для чего использует компьютерно-интенсивные информационные технологии MSA и тщательного планирования экспериментов [5, 10].

Таким образом, объектом исследования является процесс обеспечения качества продукта при реализации процессного подхода в управлении и контроле взаимосвязей между этапами производства. Предметом исследований является информационная технология оценки факторного влияния критических атрибутов процесса производства на критические атрибуты качества продукта.

Критические атрибуты выполнения важнейших (критических) процессов представляют собой результаты мониторинга критических параметров процесса производства (*CPPs*) – *X* и критических атрибутов качества продукта (*CQAs*) – *Y*. В общем случае случайные значения данных параметров (случайных величин) сформированы в массивы (таблицы) временных данных компьютерного формата от аппаратно-программных средств производства.

В терминах MSA компоненты $X - x_i$ ($i = 1,...,m$) являются предикторами, экзогенными или объясняющими переменными и рассматриваются как значения многомерной случайной величины *X*. $X^j = \{x_1,...,x_m\}$ – *m*-мерный вектор *j*-го ($j = 1,2,...,n$) наблюдения *m* критических параметров процесса или *j*-е значение случайной многомерной величины $X \supset \{X^j\}$. В свою очередь, компоненты *Y* являются эндогенными, зависимыми переменными – многомерными откликами процесса на воздействия *X*. Наблюдения $Y^j = \{y_1,...,y_k\}$ (*k* – число составляющих критических атрибутов качества продукта) однозначно соответствуют наблюдениям $X^j = \{x_1,...,x_m\}$ и составляют массив $Y \supset \{Y^j\}$. Категория $X \to Y$ описывается функцией качества процесса производства: $CQAs = F(CPPs)$.

Для достижения цели исследования, т. е. определения особенностей и характеристик изменчивости критических атрибутов качества производимого продукта *CQAs* (*Y*) от изменчивости критических параметров процесса *CPPs* (*X*) необходимо решить следующие задачи:

1. Для проведения многомерного статистического анализа массивы данных *X* и *Y* компьютерного формата должны пройти этап предварительной обработки в виде статистических процедур разведочного анализа данных (exploratory data analysis), кластеризации, факторизации и интерпретации полученных результатов.

2. Для определения и характеристики каузальных отношений $X \to Y$ в информационном пространстве переменных *X* и *Y* использовать математический аппарат многомерной ординации при детерминированном и нечётком формировании кластеров, а также процедуру построения деревьев регрессии с многомерным откликом для оценивания состава предикторов и степени их влияния на изменчивость численных значений компонентов отклика.

Процедуры 1 и 2 формулируются как задача статистического исследования структуры и иерархии многомерных данных наблюдений *X* и *Y*, а также качественной и количественной мер отношений между ними.

На практике наблюдения экзогенных переменных *X* не являются независимыми, а представляют собой группы связанных *CPPs*. Эти группы образуют кластеры с интерпретируемой зависимостью переменных, для которых представляет профильный интерес оценки влияния на многомерные отклики процесса *Y* (*CQAs*).

3. Обзор литературы. Необходимость быстрого реагирования на изменение рыночной конъюнктуры и возросшие возможности технологических средств требуют увеличения числа данных компьютерного мониторинга показателей качества процесса производства и продукта [2, 3]. При этом важная роль отводится ранним этапам проектирования, в которых закладывается базис качества продукта и определяется значительная часть его стоимости [11]. Данная стадия разработок характеризуется взаимодействием разных информационных технологий и интенсивным применением методов моделирования для улучшения качества продукции [3, 12].

В статье [13] представлен обзор литературы по проектам экспериментов для определения методологии потенциальных улучшений при моделировании путём сокращения вычислительной стоимости экспериментов и количества данных. В этом контексте применение методов выбора «наилучшего» подмножества параметров (VS, variable selection) для контроля качества выступает в качестве перспективного направления исследований [14].

Математическим аппаратом исследований параметрического качества, который рекомендуют мировые стандарты (ICH Q8, ISO 22000 HACCP [10, 15]), являются методы MSA и MSPC (Multivariate Statistical Process Control). Так, в обзоре [14] приведен анализ современного состояния интеграции VS с MSPC. При этом основной целью VS-методов заявлено определение подмножества переменных, которое несет большую часть соответствующей информации, содержащейся в полном наборе данных.

Методы MSA активно используют компьютерно-интенсивных информационные технологии [16] для уменьшения неопределенности количественного анализа за счет увеличения числа переменных в эксперименте [17]. Данные методы также применяют для определения спецификации и обоснования допусков параметров процесса (DP, design parameter) на основе анализа приемлемого изменении характеристик и чувствительности функции качества к параметрам проектирования [18].

Высокие требования к качеству продукции и усиление конкуренции – приводят к необходимости обеспечения качества продукции на стадии проектных разработок QbD [5, 10] при процессной декомпозиции функции качества (QFD) [4]. Адекватным инструментом QbD являются аналитические методы PAT [9], в основе которых – информационные технологии MSA. На этапе проектирования методы PAT сфокусированы на исследовании многофакторных связей между *CPPs* и *CQAs*, параметрами среды и их влиянием на качество конечного продукта [9, 19].

Таким образом, методы MSA создают адекватную аргументированную основу для статистической обработки массивов значений и обоснования допусков (DP, design parameter) параметров проектирования [18]. Критические *CPPs* и *CMAs* определяются на основе анализа изменений значений функции качества $CQAs = F(CPP_i)$ и их чувствительности к вариабельности аргументов. Анализ актуальных публикаций позволяет акцентировать следующие особенности поставленной задачи:

– разработка требований к методам технологического приближения к заданным значениям атрибутов *CQAs*, как некоторой функции качества от значений *CPPs*, производится на этапе первоначального проектирования процесса производства;

– гарантированное и сертифицированное качество продукта на иерархическом уровне первоначального проектирования процесса производства достигается информационными технологиями компьютерно-интенсивного многомерного статистического анализа многофакторных связей критических атрибутов качества.

4. Теоретические основы и порядок решения задачи. В общем случае зависимость объясняющих переменных (*X*) при их влиянии на многомерный отклик процесса (*Y*) обуславливает разбиение наблюдений этих многомерных параметров на однородные группы объектов, называемые кластерами или классами. Каждые из этих объектов характеризуются внутренним сходством, однако имеют детерминированные или нечёткие отличия друг от друга.

Главной задачей кластеризации является определение меры однородности объектов (наблюдений) [20]. По этому признаку различают детерминированный (экстремальный) кластерный анализ и интерпретацию компонентов наблюдений как нечетких объектов (fuzzy sets). В основе этих методов лежит анализ расстояний между всеми возможными парами объектов в пространстве наблюдаемых признаков. Такой анализ не вносит искажений в исходное взаимное положение точек $X^j = \{x_1, ..., x_m\}$ в многомерном пространстве возможных значений и обеспечивает наглядную интерпретацию зависимости наблюдений объясняющих переменных X^j при их влиянии на многомерный отклик процесса Y^j.

Детерминированные принципы кластерного анализа предполагают, что выделяемые группы представляют собой совокупности, принадлежащие только к одному кластеру. При этом принадлежность одних и тех же наблюдений или групп разным кластерам трактуется как неточность соответствующего алгоритма или большая доля случайных факторов в разбросе значений.

Полагаем, что каждый из *n* объектов $X^j = \{x_1, ..., x_m\}$ ($j = 1, 2, ..., n$) задается как точка $X_j = \{x_{j1}, x_{j2}, \ldots, x_{js}\}$ в *s*-мерном пространстве признаков $\{X_1, X_2, \ldots, X_s\}$. Совокупность этих точек можно трактовать как выборку объема *n* из многомерной генеральной совокупности $X = \{X_1, X_2, \ldots, X_s\}$.

Пусть $d(X_i, X_j) = d_{ij}$ – некоторая мера близости между каждой парой классифицируемых объектов X_i и X_j $(i = 1, 2, ..., n; j = 1, 2, ..., n)$. В качестве такой использоваться следующие меры: евклидово или манхэттенское расстояние, коэффициент корреляции Пирсона, расстояние χ^2 и другие. В случае зависимых признаков и их различной значимости при классификации объектов за меру однородности объектов принимают расстояние Махалонобиса: $d_{ij} = \sqrt{(X_i - X_j)\Lambda\Sigma^{-1}(X_i - X_j)^T}$, где $\Lambda \in R^{s \times s}$ – симметричная (чаще всего диагональная) неотрицательно определенная матрица «весовых» коэффициентов признаков, Σ – ковариационная матрица генеральной совокупности, из которой извлекаются объекты. Частными случаями меры d_{ij} являются:

– евклидово расстояние: $d_{ij} = \sqrt{(X_i - X_j)(X_i - X_j)^T} = \sum_{k=1}^{s}(x_{ik} - x_{jk})^2$, которое используется, если генеральная совокупность распределена по многомерному нормальному закону с ковариационной матрицей $\Sigma = \sigma^2 I$, а признаки однородны по своему физическому смыслу и имеют одинаковый «вес» при образовании кластеров.

– взвешенное евклидово расстояние: $d_{ij} = \sqrt{(X_i - X_j)\Lambda(X_i - X_j)^T}$;

– хеммингово расстояние (block distance): $d_{ij} = \sum_{k=1}^{s}|x_{ik} - x_{jk}|$.

Таким образом, процедуры выбора метрики расстояний между классами объектов d_{ij} и формирования матрицы расстояний между *n* объектами $D = d_{ij} \in R^{s \times s}$ определяют разбиение *n* объектов на *p* кластеров: $d_{ij} = p(X_i, X_j)$.

При образовании групп нечетких объектов X_j, множество *X* является нечетким, если для заданной на интервале [0, 1] функции принадлежности (membership function) $\mu_x(X_j)$, значение $\mu_x(X_j)$=0 означает полную несовместимость, т. е. $X_j \notin X$, а $\mu_x(X_j)$=1 означает полную принадлежность (или $X_j \in X$) наблюдений X_j к множеству *X*. Функция $\mu_{jk}(X_j)$ задает в масштабе от 0 до 1 степень принадлежности каждого объекта X_j к каждому выделяемому кластеру *K*. За меру расстояний при нечётком разбиении на кластеры принимается метрика $d_{jk} = \sum_i (x_{ji} - v_{ki})$ – расстояние между *j*-ым объектом с набором признаков x_{ji} и центрами тяжести v_{ki} каждого *k* кластера $k = 1, 2, ..., p$. В общем, задача нечеткой кластеризации объектов X_j состоит в нахождении матрицы μ, которая минимизирует критерий:

$$R_{pm} = \sum_{j=1}^{n}\sum_{k=1}^{p}\mu_{jk}^m d_{jk} = \min .$$

Степенной вес *m* определяет уровень нечеткости получаемых кластеров: чем больше *m*, тем более нечетким является разбиение. Рекомендуемый диапазон варьирования *m* задаётся в интервале {1.2, 2}. Количество кластеров определяется в результате разведочного анализа данных при реализации алгоритмов с различными метриками расстояний.

Важно отметить, что задача проверки адекватности кластеризации является не решённой теоретически. Распространённым практическим критерием адекватности является сравнение принадлежности групп классам, полученным разными методами кластеризации или априорное знание принадлежности объектов к соответствующим классам. В статье использовался метод объединения в кластеры с перебором различных комбинаций числа групп, метрик дистанций и выбором результата кластеризации по 30 индексам качества. Данный метод реализован в программном пакете NbClust языка R. Принципы кластеризации и анализа результатов определяют следующий порядок решения задачи:

1. Проведение разведочного анализа данных предикторов *X*: определение корреляции переменных и оценка тенденции кластеризации.

2. Формализация задачи, т.е. проведение кластеризации поясняющих переменных *X* по выбранным статистическим методам. На данном шаге массив наблюдений *X* кластеризуется по различным детерминированным метрикам с определением варианта по 30 индексам. Результаты визуализируются.

3. Отождествление полученных кластеров переменных *X* с факторами влияния на переменные *Y*. На данном этапе обработки значений критических атрибутов *X* и *Y* выполняется важная процедура трансформации компьютерного формата данных в структурированные таблицы временных данных, подготовленные к факторному анализу.

4. Проведение и анализ результатов прямой и непрямой ординации откликов *Y* и сравнение их с результатами влияния факторов *X* на многомерные отклики *Y*. Проводится идентификация групп наблюдений по принципу нечёткой классификации и сравнение её с детерминированными методами кластеризации.

5. Анализ результатов сравнения по п.4 и уточнение формализации задачи для проведения операций по п.2 для детализации выводов с практическими и исследовательскими целями.

В общем случае, приведенный алгоритм соответствует общепринятой схеме установления физических закономерностей: сбор экспериментальных данных, организация их в виде таблиц и поиск такого способа обработки, который позволил бы обнаружить в исходных данных новые знания об анализируемом процессе. Вычисления и оценки произведены с помощью объектно-ориентированного языка программирования R.

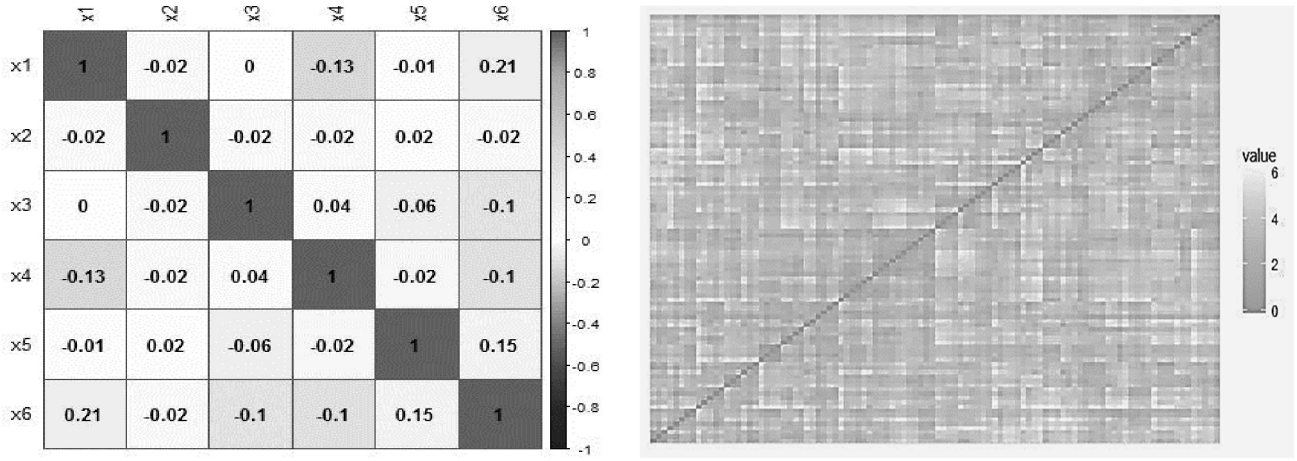
Приведенные результаты соответствуют статистическим оценкам, которые проводились с многомерными массивами временных данных предикторов *X* и откликов *Y*, заданных в виде равномерно распределённых случайных величин в пределах допусков.

5. Анализ результатов решения задачи. Разведочный анализ данных использовался для описательной статистики свойств выборок и включал: обработку и систематизацию эмпирических данных, представление многомерных данных в форме таблиц, а также количественные характеристики основных статистических показателей.

Предварительные выводы о свойствах предикторов можно сделать по корреляционной матрице и визуализации тенденций образования кластеров VAT (Visual Assessment of cluster Tendency), как показано на рис. 2*a*,*b*. Результаты оценки веса или "важности" предикторов *X* (variable importance) и объединения предикторов *X* в кластеры, показаны на рис. 3*a*,*b*.

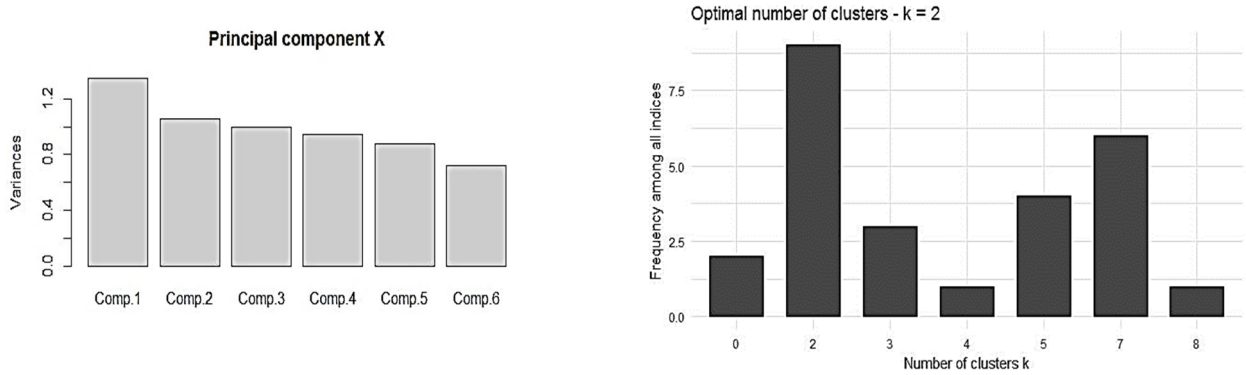
На рис. 3*a*,*b*: principle component (главная компонента), variance (дисперсия), frequency among all indices (частота среди всех индексов качества кластеризации), number of clusters (количество кластеров), optimal number of clusters (оптимальное количество кластеров).

Благодаря факторизации массива объясняющих переменных *X* и использованию ординационного анализа произведено распределение значений массива многомерных откликов *Y* по объектам факторов на плоскости главных компонент PCA (PCA, principle components analysis).



a *b*

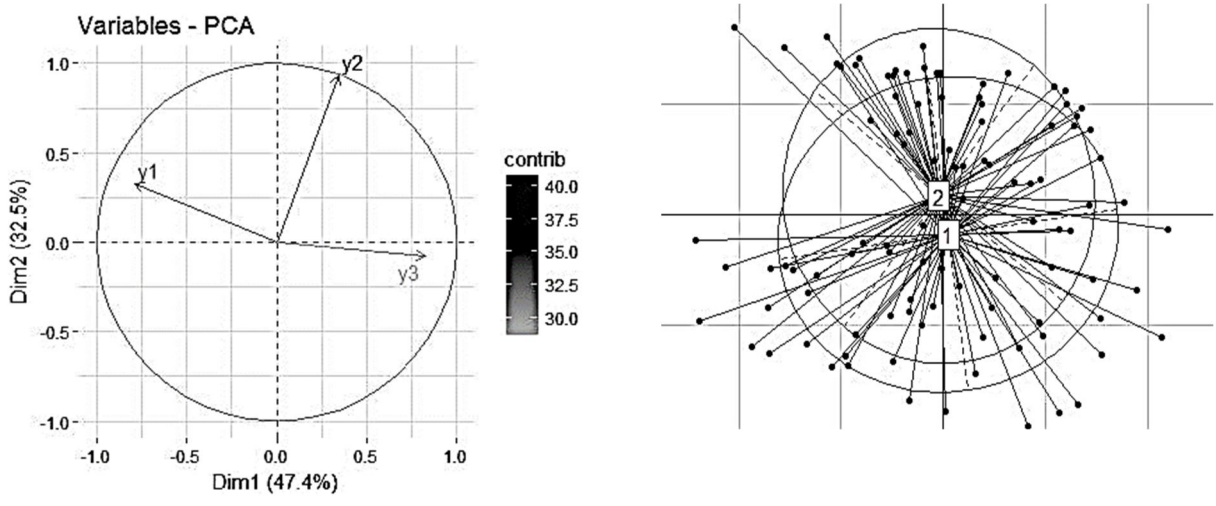
Рис. 2. *a)* Результат оценки корреляции предикторов; *b)* и визуализация тенденций образования кластеров



a *b*

Рис. 3. *a)* Результат оценки веса компонент предикторов X^j;
b) результат определения оптимального числа кластеров по оценкам различных индексов

Внутренние особенности массива Y показаны на рис. 4*a* в виде непрямой ординации составляющих на плоскости главных компонент; распределение значений Y^j по классам объектов (факторов) массива X в виде прямой ординации изображено на рис. 4*b*.



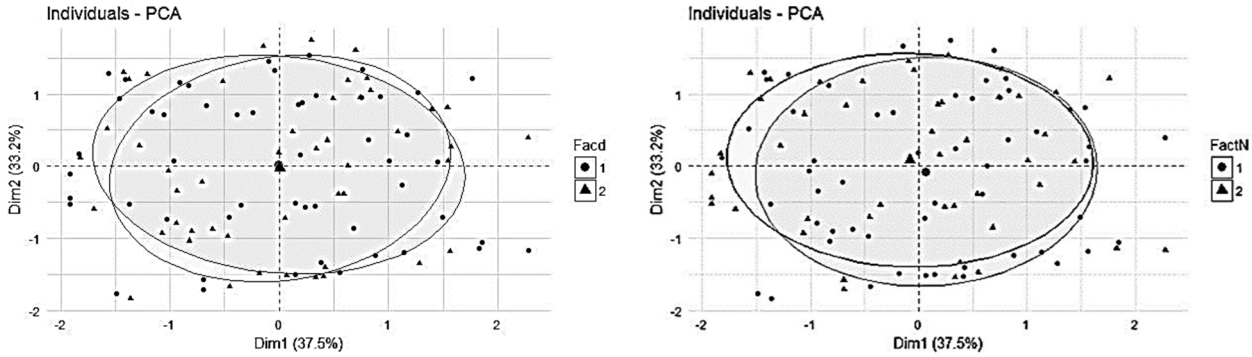
a *b*

Рис. 4. Диаграммы непрямой – *a* и прямой – *b* ординации массива откликов Y на плоскости главных компонент PCA

На рис. 4 *a*,*b*: Dim1, 2 (главные компоненты на рис.4*a* и аналогично на рис.4*b*), contribution (вклад каждой переменной (variable) y_j в формирование каждой главной компоненты (PCA) Dim.). Рисунок 4*b* иллюстрирует группировку наблюдений откликов Y^j по факторным признакам на фоне объясняющих составляющих x_m (пунктирные линии) связанных с собственными векторами массива X (ординационный триплот).

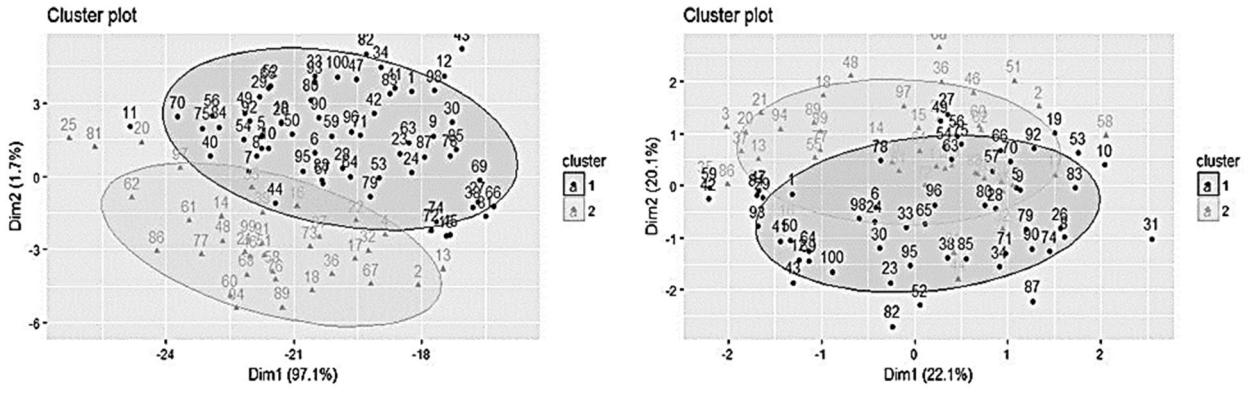
Важно, что предложенная трансформация кластеров объясняющих переменных X в факторы позволяет провести прямую (constrained) ординацию как при детерминированной, так и при нечёткой кластеризации наблюдений X^j . Процедура нечёткого формирования объектов даёт более гибкую интерпретацию принадлежности наблюдений Y^j к факторам, что на практике раскрывает скрытые влияния критических параметров процесса производства (*CPPs*) на критические атрибуты качества продукта (*CQAs*) .

Так приведём прямые ординациионные диаграммы откликов Y^j с факторизацией по детерминированному методу k медоидов или PAM (Partitioning Around Medoids) и с факторизацией по нечёткой логике с весовым критерием, рис. 5. Каждому из этих вариантов соответствует кластеризация объясняющих переменных X обоими способами, рис. 6. В итоге рис. 5 визуализирует двумерные диаграммы (ординационный биплот) распределения наблюдений Y по двум видам кластеров (факторов, Facd/FactN), которые формируются с предварительным приведением пространства откликов Y к двум главным компонентам (Dim1, 2).



a *b*

Рис. 5. Диаграммы прямой детерминированной – *a* и прямой нечёткой – *b* ординации массива откликов Y на плоскости главных компонент (individuals) PCA – Dim1, 2.



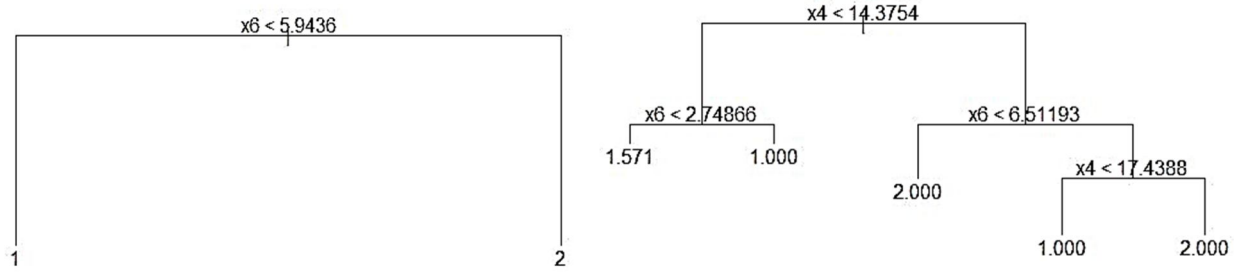
a *b*

Рис. 6. Диаграммы (cluster plot) непрямой детерминированной – *a* и непрямой нечёткой – *b* кластеризации массива объясняющих переменных X на плоскости главных компонент PCA

Из рис. 5 видно, что классы различаются слабо. Критерием разделения является показатель observation: 0.006107815, полученный по тесту Monte-Carlo. Поскольку этот показатель меньше 0.5, – различия между классами определены как нечёткие.

Далее целесообразно показать, какие именно компоненты объясняющих переменных X доминируют при кластеризации наблюдений X^j . При введённых факторных переменных в однородные таблицы значений объясняющих переменных X этот анализ производится с помощью построения так называемых "деревьев классификации", или «деревьев решений» ("classification trees", или "decision trees"), рис.7.

Как видно из рис. 7, результаты различаются для кластеров полученных детерминированным (рис. 7*a*) и нечётким методом (рис. 7*b*).



a *b*

Рис. 7. Деревья классификации при детерминированной – *a* и нечёткой – *b* кластеризации массива объясняющих переменных X

В обоих случаях "листьями" полученного дерева являются кластеры объектов, образованные по принципу минимизации различий между точками в многомерном пространстве объясняющих переменных X в пределах каждой совокупности.

Анализ может быть конкретизирован при построении деревьев регрессии массива объясняющих переменных X с многомерными откликами Y (MRT, Multivariate Regression Trees), когда решается задача оценивания влияния составляющих объясняющих переменных $\{x_1,...,x_m\}$ на совокупную изменчивость составляющих многомерного отклика $\{y_1,...,y_k\}$, как показано на рис. 8.

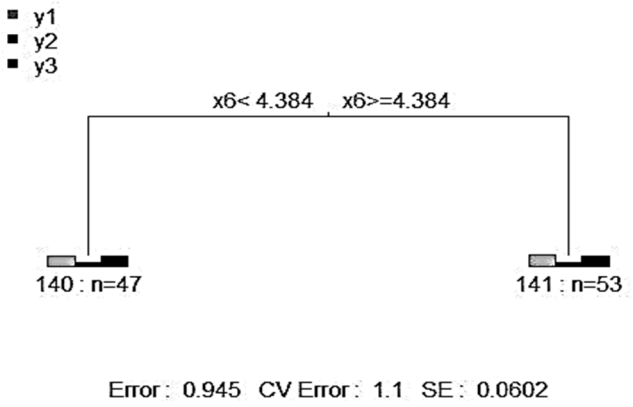


Рис. 8. Дерево классификации с многомерным откликом $\{y_1,...,y_k\}$ для кластеров массива экзогенных переменных X

Из рис. 7 и из рис. 8 виден характер и степень влияния составляющих объясняющих переменных X на многомерный отклик процесса Y.

Таким образом, многомерные методы статистического анализа позволяют выполнить кластеризацию массива наблюдаемых объясняющих переменных X, что увеличивает число степеней свободы при анализе их влияния на наблюдения многомерного отклика процесса Y и позволяет совмещать технологии статистического оценивания для детальной практической интерпретации результатов.

6. Обсуждение

В статье представлены результаты исследования на этапе проектирования процесса производства, в ходе которого проводился многомерный статистический анализ причинно-следственных связей между изменчивостью компонент отклика процесса и действием определенных факторов, отождествляемых с технологическими параметрами процесса.

Факторы получены как результат анализа группирования наблюдений массива объясняющих переменных в кластеры для разных метрик оценивания разброса их значений в многомерном пространстве. Исследование содержит элементы разведочного, описательного и аналитического статистического анализа.

Оценка тенденции кластеризации массива X (рис. 2*b*) и "важности" предикторов $\{x_1,...,x_m\}$ (variable importance) (рис.3*a*) с учётом возможной мультиколлинеарности компонент (рис. 2*a*) анализировались с помощью метода главных компонент PCA и визуализации тенденций образования кластеров VAT. Метод PCA даёт значения относительного ожидаемого вклада каждой компоненты x_m (Comp.m) в общую вариацию данных, как показывают данные в табл. 1.

Отметим, что все компоненты (Comp.n) многомерного массива X характеризуются практически равным вкладом (proportions of variance 16.6%) в разброс данных. Тенденция кластеризации по статистике Хопкинса (The Hopkins statistic, 0.5087) указывает на умеренную склонность к образованию кластеров при случайном разбросе данных.

Показатели вклада компонент x_m в общую вариацию значений массива X **Табл. 1**

Comp. / Параметры	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Стандартное отклонение	1.1597	1.0287	0.9966	0.9712	0.9378	0.8486
Пропорциональный вклад	0.2264	0.1781	0.1672	0.1588	0.1480	0.1212
Кумулятивная пропорция	0.2264	0.4046	0.5718	0.7306	0.8787	1.0000

Отметим, что все компоненты многомерного массива X характеризуются практически равным вкладом (proportions of variance 16.6%) в разброс данных. Тенденция кластеризации по статистике Хопкинса (The Hopkins statistic, 0.5087) указывает на умеренную склонность к образованию кластеров при случайном разбросе данных. В основу критерия качества кластеризации положена метрическая гипотеза компактности Загоруйко. По этому критерию получены результаты оценки качества кластеризации для массива многомерных данных X (рис. 3б), как показано в табл. 2.

Кластерные расстояния по тесту Загоруйко **Табл. 2**

Средние кластерные расстояния / № кластера	Cl. dist. 1	Cl. dist. 2
1	3.044665	3.623451
2	3.623451	3.009617

В табл. 2 размещены средние кластерные расстояния (Cl. dist., cluster distances): внутрикластерные (диагональные элементы таблицы, intra-cluster distances), которые меньше, чем межкластерные расстояния (недиагональные элементы таблицы, cross-cluster distance), что свидетельствует о допустимости разбиения пространства наблюдений массива X на два кластера.

Операция трансформации однородной таблицы наблюдений откликов Y в таблицу с факторными переменными была проведена после исследования особенностей этого массива (рис. 4ab). Результаты проекции в ортогональной системе координат главных компонент

(Dim.n) составляющих массива многомерного отклика Y (непрямая ординация, рис. 4а) показаны в табл. 3.

Из табл. 3 видно, что составляющие массива откликов Y: $\{y_1,...,y_k\}$ по степени своего вклада в компоненты разложения Dim1, Dim2, Dim3 факторы не образуют и имеют слабую корреляцию.

Вклад откликов $\{y_1,...,y_k\}$ в формирование осей главных компонент (Dim.*n*) **Табл. 3**

Dim / Y	Dim. 1	Dim. 2	Dim. 3
y1	43.56287	10.7237730	45.71336
y2	8.30943	88.6417708	3.04880
y3	48.12770	0.6344562	51.2378

Прямая ординация массива откликов (рис. 4б) показана на фоне составляющих $\{x_1,...,x_m\}$ (пунктирные линии) и связывает внутренние закономерности откликов Y^j с факторизованными воздействиями критических параметров процесса X^j. Попадание одноименных откликов Y^j в оба факторных объекта свидетельствует о слабой корреляции составляющих $\{y_1,...,y_k\}$ и преобладании случайных влияний компонент $\{x_1,...,x_m\}$ на изменчивость составляющих откликов Y^j. Этот факт обуславливает необходимость в дальнейшем провести исследования компонентного влияния предикторов $\{x_1,...,x_m\}$ на изменчивость составляющих многомерной зависимой случайной величины Y^j.

Как видно из рис. 5 при детерминированной (*a*) и нечёткой (*b*) кластеризации в сформированные объекты попадают разные наблюдения откликов Y^j. Очевидно, что при попадании наблюдения Y^j в один и тот же кластер для двух вариантов группирования будет свидетельствовать о уверенном влиянии данной определённой совокупности предикторов X^j.

Соответствующее ординации на рис. 5 разбиение массива наблюдений предикторов X на объекты по обоим метрикам логик показано на рис. 6. Видно, что изменение метода кластеризации существенно сказывается на ассоциации наблюдений массива X. Эта особенность затем проявляется при оценивании составляющих объясняющих переменных X, которые являются разделяющими при кластеризации наблюдений, методом построения деревьев классификации – рис.7. По результатам визуализации можно предположить, что детерминированные методы дают более грубые оценки (рис. 7a) по сравнению с нечёткой интерпретацией (рис. 7b). Сближение результатов можно ожидать при большей тенденции объясняющих переменных X к кластеризации, а также меньшей доли случайной составляющей в разбросе этих значений.

На заключительном этапе анализа было сформировано MRT массива экзогенных переменных X с многомерным откликом Y, как показано на рис. 8. Из рис. 8 видно, что разделяющей переменной является компонента *x6*, которая образует равнозначные кластеры с практически равным числом наблюдений (47 и 53) при отсутствии явного доминирования составляющих y_k. Примечательно, что результат MRT практически совпадает с результатом построения дерева классификации при детерминированной кластеризации массива объясняющих переменных X (рис. 7a). Это является следствием того, что метод MRT использует детерминированные методы кластеризации при построении деревьев решений.

Обсуждение показывает, что на этапе проектных разработок процесса производства к экспериментальным временным данным компьютерного формата принципиально необходимо применять информационные технологии MSA для исследования, анализа и интерпретации

потенциально возможных закономерностей и зависимостей между отдельными компонентами и объектами этих многомерных данных.

6. Выводы

Методы многомерного статистического анализа являются удобным и адекватным инструментом конструирования многомерной функции качества и исследования её критических параметров на каждом из этапов процесса производства. Эти методы целесообразно применять на этапе первичного проектирования процесса производства при проведении статистических экспериментов для обеспечения качества продукции.

Более полные и детализированные результаты исследований можно получить путём комбинирования каузальных методов многомерного статистического анализа, что позволяет оперировать численными характеристиками с адекватной практической интерпретацией влияния критических параметров процесса производства на критические атрибуты качества продукта.

Новизна приведенных результатов состоит в кластеризации временных данных критических параметров процесса производства и отождествлении их с факторами влияния на соответствующие временные данные критических атрибутов качества продукта. Данная процедура даёт существенные дополнительные степени свободы для применения методов многомерного статистического анализа с целью исследования влияния критических параметров процесса производства на критические атрибуты качества продукта.

Перспектива дальнейших исследований направлена на синтез или применение робастных методов обеспечения стандартных показателей критических атрибутов качества продукции, которые устойчивы к воздействию различных факторов среды и к вариабельности критических параметров процесса производства.

Список использованной литературы

1. Introduction and support package: Guidance on the concept and use of the process approach for management systems // ISO 9000.

2. Yukish M. A. A preliminary model of design as a sequential decision process / M. A. Yukish, S. W Millerb, T. W. Simpson // Procedia computer science. – 2015. – No.44 – P. 174-183.

3. Marijn P. Z. Physics in Design: Real-time numerical simulation integrated into the CAD environment / P. Z. Marijn, W. W. Wessel // Procedia CIRP 60. –2017. – P. 98-103.

4. Lionberger R. A. Quality by design: concepts for ANDAs / R. A. Lionberger, S. L. Lee, L. M. Lee, A. Raw, X. Yu. Lawrence // The AAPS Journal. – 2008. – Jun, 10(2). – P. 268-276.

5. Zhang L. Application of quality by design in the current drug development / L. Zhang, S. Mao // Shenyang Pharmaceutical University, Wenhua Road, Shenyang, China: Asian journal of pharmaceutical sciences. – 2017. – No.103. – P. 1-8.

6. Rao S. An Overview of Taguchi method: evolution, concept and interdisciplinary applications / S. Rao, P. Samant, A. Kadampatta, R. Shenoy // International journal of scientific & engineering research. – October 2013. – Vol. 4, Issue 10. – P. 621-626.

7. Guidance for industry. Quality systems approach to pharmaceutical CGMP regulations. – https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm070337.pdf (2018.06.20).

8. Jaiswal E. S. A Case study on quality function deployment (QFD) / E. S. Jaiswal // IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE) – Nov-Dec. – 2012. – Volume 3, Issue 6 – P. 27-35.

9. Guidance for industry PAT. A framework for innovative pharmaceutical development, manufacturing, and quality assurance. – http://www.fda.gov/cder/OPS/PAT.htm (2018.06.20).

10. Pramod K. Pharmaceutical product development: A quality by design approach / K. Pramod, M. A. Tahir, N. A. Charoo, S. H. Ansari, J. Ali // Int J Pharm Investig. – Jul-Sep 2016. – 6(3). – P. 129-138.

11. Steimera C. Model-based design process for the early phases of manufacturing system planning using SysML / C. Steimera, J. Fischerb, J. C. Auricha // 27th CIRP Design 2017. Procedia CIRP. – 2017. – P. 163-168.

12. Mohamed I. Progressive modeling: The process, the principles, and the applications / I. Mohamed // Procedia Computer Science. – 2013. – Volume 16. – P. 39-48.

13. Blondet G. Simulation data management for adaptive design of experiment. A literature review / G. Blondet, N. Boudaoud, J. Duigou // QUALITA' 2015. – Mar 2015, Nancy, France.

14. Araujo F. Variable selection methods in multivariate statistical process control: A systematic literature review / F. Araujo, P. Pimentel, F. S. Fogliatto // Department of industrial engineering, Federal University of Rio Grande do Sul, 90035-190 Porto Alegre, RS, Brazil, Computers & Industrial Engineering. – January 2018. – Volume 115. – P. 603-619.

15. Food safety management systems // ISO 22000: 2005.

16. Skold M. Computer intensive statistical methods / M. Skold // Mathematical statistics centre for mathematical sciences Lund University. – October 2005, 2nd printing August 2006. – 133 p.

17. Brereton R. G. Applied chemometrics for scientists / R. G. Brereton. – University of Bristol. – UK, John Wiley & Sons Ltd, 2007. – 379 p.

18. Göhler S. M. The Translation between Functional Requirements and Design Parameters for Robust Design / S. M. Göhler, S. Husung, T. J. Howard // 14th CIRP conference on computer aided tolerancing (CAT). Procedia College International pour la Recherche en Productique (CIRP) 43. – 2016. – P. 106-111.

19. Food and drug administration. Final report on pharmaceutical cGMPs for the 21st century – A Risk based approach. – http://www.fda.gov/cder/gmp/gmp2004/GMP_-finalreport2004.htm (2018.06.20).

20. Kogan J. Grouping multidimensional data / J. Kogan, C. Nicholas, M. Teboulle. – Springer-Verlag Berlin Heidelberg 2006. – 268 p.

Автори статті

Курченко Олег Анастасійович – кандидат технічних наук, доцент, завідувач кафедри управління інформаційною та кібернетичною безпекою, Державний університет телекомунікацій, Київ, Україна. Тел.: +380 (96) 714 82 39. E-mail: kurol@ukr.net. ORCID 0000-0002-3507-2392.

Терещенко Антон Ігорович – аспірант, кафедра управління інформаційною та кібернетичною безпекою, Державний університет телекомунікацій, Київ, Україна. Тел.: +380 (98) 817 09 42. E-mail: iter@ukr.net. ORCID: 0000-0002-4728-9652.

Authors of the article

Kurchenko Oleh Anastasiiovych – candidate of sciences (technical), head of the department of management of information and cyber security, State University of Telecommunications, Kyiv, Ukraine. Tel. +380 (96) 714 82 39. E-mail: kurol@ukr.net. ORCID 0000-0002-3507-2392.

Tereshchenko Anton Ihorovych – post-graduate student, department of management of information and cyber security, State University of Telecommunications, Kyiv, Ukraine. Tel.: +380 (98) 817 09 42. E-mail: iter@ukr.net. ORCID ID: 0000-0002-4728-9652.

Дата надходження
в редакцію: 05.03.2018 р.

Рецензент:
доктор технічних наук, професор
М. М. Степанов
Державний університет телекомунікацій, Київ