ITHEA

International Journal INFORMATION MODELS & ANALYSES

2015 Volume 4 Number 3

International Journal
INFORMATION MODELS & ANALYSES
Volume 4 / 2015, Number 3

IJ IMA is official publisher of the scientific papers of the members of the ITHEA® International Scientific Society

IJ IMA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for ITHEA International Journals are given on www.ithea.org .
The camera-ready copy of the paper should be received by ITHEA® Submission system http://ij.ithea.org .
Responsibility for papers published in IJ IMA belongs to authors.

International Journal "INFORMATION MODELS AND ANALYSES" Volume 4, Number 3, 2015

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
National Aviation University, Ukraine
Universidad Politechnika de Madrid, Spain,
Hasselt University, Belgium
Institute of Informatics Problems of the RAS, Russia,
St. Petersburg Institute of Informatics, RAS, Russia
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia,

Publisher: **ITHEA®**

Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Technical editor: **Ina Markova**

Printed in Bulgaria

ISSN 1314-6416 (printed) **ISSN 1314-6432 (Online)**

О СХОДИМОСТИ ПОСЛЕДОВАТЕЛЬНОСТЕЙ НЕЧЕТКИХ ПЕРСЕПТИВНЫХ ЭЛЕМЕНТОВ, ЗАДАННЫХ НА РАЗНЫХ ПРОСТРАНСТВАХ ВОЗМОЖНОСТЕЙ

Алексей Бычков, Евгений Иванов, Ольга Супрун

Abstract: *В статье получены критерии существования модели сходимости и расходимости с необходимостью 1 для систем конечномерных распределений последовательностей нечётких персептивных элементов, доказаны теоремы о не выполнении теоретико-возможностного аналога закона больших чисел для сходимости по возможности и сходимости с необходимостью 1.*

Keywords: *теория возможностей, нечёткая логика, сходимость нечётких персептивных элементов.*

ACM Classification Keywords: *G.3 – Probability and Statistics; I – Computing Methodologies, I.6 – Simulation and Modeling.*

Введение

В работах [Пытьев, 2000], [Бычков, 2007a], [Бычков, 2007b] для моделирования неопределенностей предлагается применять теорию возможностей. Основы этой теории заложены в [Zadeh, 1978] и монографии [Дюбуа, 1990]. В этих работах вводится понятие мер возможности, необходимости и основные аксиомы построения пространства возможностей.

Из монографии [Пытьев, 2000] известен теоретико-возможностный аналог закона больших чисел для сходимости распределений, но для других основных видов сходимости в теории возможностей (по возможности и с необходимостью 1) не было известно, имеет ли место такой аналог.

В этой статье исследуем свойства сходимости последовательностей нечетких персептивных элементов, заданных на разных возможностных пространствах и применим полученные результаты к исследованию теоретико-возможностного аналога закона больших чисел для сходимости по возможности и с необходимостью 1.

Основной результат

Пусть X – не пустое множество (пространство элементарных событий), $\mathbf{A}$ – класс подмножеств X, который содержит $\varnothing$ и X (множество составных событий), $\beta(X)$ – булеан множества X.

Определение 1. Полностью аддитивной мерой возможности на классе множеств $\mathbf{A}$ называется функция $P: \mathbf{A} \to [0,1]$, которая удовлетворяет условию

$$P\left(\bigcup_{t\in T} A_t\right) = \sup_{t\in T} P(A_t)$$

для каждого семейства $\{A_t \mid t \in T\}$ множеств из класса $\mathbf{A}$ такого, что $\bigcup_{t\in T} A_t \in \mathbf{A}$.

Мера возможности P называется нормированной, если $P(X)=1$ и $P(\varnothing)=0$.

В дальнейшем, если не сказано иначе, меры возможности будут считаться нормированными и полностью аддитивными.

Определение 2. P -моделью теории возможностей называется тройка $(X, \mathbf{A}, P)$, где $\{0, X\} \subseteq \mathbf{A} \subseteq 2^X$, P -мера возможности на классе множеств $\mathbf{A}$. P -модель мы также будем называть пространством возможностей.

Для дальнейшего нам понадобится техника продолжения меры возможности с одного класса множеств на более широкий класс множеств. Проблема продолжения меры возможности рассматривалась многими авторами [Boyel,1995], [Wang, 1992], [Бычков, 2007a], [Пытьев, 2000]. Мы будем использовать следующий вариант теоремы о продолжении [Wang, 1992].

Определение 3. Функция $P^*(A) = \inf\left\{\sup_{t\in T} P(A_t) \mid \bigcup_{t\in T} A_t \supseteq A\right\}$, где нижняя грань берётся по семействам множеств $(A_t)_{t\in T}$ из класса $\mathbf{A}$, которые покрывают множество A, называется внешней мерой возможности, соответствующей функции $P: \mathbf{A} \to [0,1]$.

Теорема (О продолжении меры возможности).

1) Функцию P, определённую на классе множеств $\mathbf{A} \subseteq 2^X$ можно продолжить до меры возможности на $\beta(X)$ тогда и только тогда, когда для произвольного семейства множеств $(A_j)_{j\in J}$, $A_j \in \mathbf{A}$ и множества $A \in \mathbf{A}$ выполняется импликация

$$A \subseteq \bigcup_{j\in J} A_j \Rightarrow P(A) \le \sup_{j\in J} P(A_j).$$

2) Если P имеет некоторое продолжение $\overline{P}$ до меры возможности на $\beta(X)$, то P^* является продолжением P до меры возможности на $\beta(X)$ и $\forall A \subseteq X\, \overline{P}(A) \le P^*(A)$.

Следствие. Если класс множеств $\mathbf{A}$ замкнут относительно конечных пересечений и P-мера возможности, то P^* является её (наибольшим) продолжением до меры возможности на $\beta(X)$.

Пространство возможностей $(X, \mathbf{A}, P)$ назовём регулярным, если $\mathbf{A} = \beta(X)$ и мера возможности $P : \beta(X) \to L$ является полностью аддитивной и нормированной.

Пусть задано метрическое пространство $\mathbf{M} = (M, d)$.

Определение 4. Нечётким персептивным элементом ξ на пространстве возможностей $(X, \mathbf{A}, P)$ называется $\mathbf{A}$-измеримая тотальная функция $\xi : X \to \mathbf{M}$.

Определение 5. В случае, когда $\mathbf{A} = \beta(X)$, распределением нечёткого персептивного элемента ξ называется функция $f_\xi(y) = P\{\xi = y\}$, $y \in \mathbf{M}$.

Введем следующие обозначения:

$M^+ = \bigcup_{k=1}^{\infty} M^k$ – множество кортежей из элементов M (слов в алфавите M);

M^ω – множество последовательностей элементов M (ω-слов в алфавите M);

$a * b$ – конкатенация кортежей (слов) a и b, где $a \in M^+$, $b \in M^+ \cup M^\omega$;

$a \prec b$ – отношение “быть строгим префиксом”, т.е. $\exists c \in M^+ \cup M^\omega \;\; a * c = b$;

$\beta_+(M) = \bigcup_{n\geq 1} \beta(M^n)$ – множество слов фиксированной конечной длины;

$\beta_\infty(M) = \beta_+(M) \cup \beta(M^\omega)$.

Для пар множеств $A \in \beta_+(M)$, $B \in \beta_\infty(M)$ введём обозначения:

$\mathrm{len}(A) = n$, если $A \subseteq M^n, A \neq \varnothing$;

$A * B = \{a * b \mid a \in A, b \in B\} \in \beta_\infty(M)$ – конкатенация всех пар элементов;

$\mathrm{Pref}_n(B) = \{a \in M^n \mid \exists b\ a * b \in B\}$ – множество префиксов длины n;

$\mathrm{Suff}(B) = \{w \mid \exists u \in M^*\ u * w \in A\}$ – множество суффиксов.

Определим следующие множества:

1) FD_M – множество полностью аддитивных нормированных мер возможности на $\beta(M)$. Его элементы будем называть (нечеткими) распределениями;

2) FD_M^ω – множество последовательностей распределений (элементов FD_M);

3) FS_M^ω – множество полностью аддитивных нормированных мер возможности на $\beta(M^\omega)$. Его элементы будем называть (нечеткими) распределениями последовательности;

4) FS_M^+ – множество полностью аддитивных нормированных мер возможности Q_+ на $\beta_+(M)$, которые удовлетворяют условию $Q_+(Y) = Q_+(Y * M)$ для всех $Y \in \beta_+(M)$. Его элементы будем называть системами конечномерных распределений последовательности.

Определим тотальный оператор $Fin : FS_M^\omega \to FS_M^+$ равенством

$$(Fin(Q))(Y) = Q(Y * M^\omega),\ Y \in \beta_+(M).$$

Введем на множестве M^ω топологию: открытые множества имеют вид $W * M^\omega$, где $W \subseteq M^+$ (топология произведения). Множества из $BT_M^0 = \{u * M^\omega \mid u \in M^*\}$ будем считать (открытыми) шарами. Класс открытых множеств будем обозначать как $BT_M = \{U * M^\omega, U \in \beta_+(M)\}$.

Утверждение 1. Выполняются следующие свойства:

1) класс BT_M замкнут относительно конечных объединений и пересечений;

2) тотальная функция $Q_\omega^0 : BT_M \to L$, определенная равенством $Q_\omega^0(U * M^\omega) = Q_+(U)$, является мерой возможности на BT_M. Будем ее обозначать как $Inf(Q_+)$;

3) для $Q_+ \in FS_M^+$ и $Q_\omega \in FS_M^\omega$, $Fin(Q_\omega) = Q_+$ тогда и только тогда, когда $Q_\omega|_{BT_M} = Inf(Q_+)$.

Доказательство.

1) Следует из равенства $U_1 * M^\omega \circ U_2 * M^\omega = (U_1 \circ U_2) * M^\omega$, где $\circ$ обозначает $\cup$ или $\cap$.

2) Пусть $U * M^\omega = U' * M^\omega$ для некоторых $U \in M^k, U' \in M^n$, считаем $k \le n$. Тогда $U * M^{n-k} = \text{Pref}_n(U * M^\omega) = \text{Pref}_n(U' * M^\omega) = U'$ и, следовательно: $Q_+(U) = Q_+(U')$.

Поэтому функция Q_ω^0 является корректно определенной. Рассмотрим семейство множеств:

$$Y_t = U_t * M^\omega \in BT_M, t \in T \text{ такое, что } Y = \bigcup_{t \in T} Y_t = U * M^\omega \in BT_M.$$

Пусть $n = \text{len}(U)$. Тогда $U = \bigcup_i U_i * M^{n-len(U_i)}$, где объединение по таким i, что $\text{len}(U_i) \le n$. И соответственно:

$$Q_\omega^0(Y) = Q_+(U) = \sup_i Q_+(U_i * M^{n-len(U_i)}) = \sup_i Q_+(U_i) = \sup_{t \in T} Q_+(U_t) = \sup_{t \in T} Q_\omega^0(Y_t).$$

Следовательно, Q_ω^0 является полностью аддитивной мерой возможности на BT_M.

3) Необходимость. Пусть $Fin(Q_\omega) = Q_+$. Тогда $Q_\omega(U * M^\omega) = Q_+(U) = Inf(Q_+)(U * M^\omega)$.

Достаточность. Пусть $Q_\omega|_{BT_M} = Inf(Q_+)$. Тогда $Q_+(U) = Inf(Q_+)(U * M^\omega) = Q_\omega(U * M^\omega)$, откуда, по определению, $Fin(Q_\omega) = Q_+$.

Утверждение доказано.

Введём такое обозначение: если $\xi_n : X \to M$ – последовательность нечетких персептивных элементов, то нечеткий персептивный элемент $\xi_{(n)} : X \to M^{\omega}$ определяется как $\xi_{(n)}(x) = (\xi_1(x), \xi_2(x), \ldots)$.

Утверждение 2. Пусть $(X, 2^X, P)$ – пространство возможностей, $\xi_n : X \to M$ – последовательность нечетких персептивных элементов, $\xi : X \to M$ – нечёткий персептивный элемент. Тогда $P_{\xi} \in FD_M$, $P_{\xi_{(n)}} \in FS_M^{\omega}$ и $Fin(P_{\xi_{(n)}}) \in FS_M^{+}$.

Доказательство. Очевидно.

Определение 6. Пара (PS, ξ), где PS – регулярное пространство возможностей $(X, 2^X, P)$, $\xi : X \to M$ – нечеткий персептивный элемент, называется моделью распределения $Q \in FD_M$, если $Q \equiv P_{\xi}$.

Аналогично пара $(PS, \xi_{(n)})$, $\xi_{(n)} : X \to M^{\omega}$ называется моделью распределения последовательности $Q_{\omega} \in FS_M^{\omega}$.

Моделью последовательности распределений $Q_{(n)} \in FD_M^{\omega}$ называется пара $(PS, \xi_{(n)})$, $\xi_{(n)} : X \to M^{\omega}$, такая, что $\forall n \in \mathbf{N} \quad Q_n \equiv P_{\xi_n}$.

Моделью системы конечномерных распределений последовательности $Q_{+} \in FS_M^{+}$ называется пара $(PS, \xi_{(n)})$, такая, что $Q_{+} = Fin(P_{\xi_{(n)}})$.

Определение 7. Моделью сходимости с необходимостью 1 системы конечномерных распределений последовательности Q_{+} называется такая модель $(PS, \xi_{(n)})$ системы распределений Q_{+}, в которой $\xi_{(n)}$ сходится с необходимостью 1.

Аналогично определяться понятия модели расходимости с необходимостью 1 и модели сходимости (расходимости) с положительной необходимостью.

Пусть $(X, \mathbf{A}, P)$ – пространство возможностей, в котором класс множеств $\mathbf{A}$ замкнут относительно конечных пересечений, а P – нормированная мера возможности. Пусть P^* – внешняя мера возможности, соответствующая P.

Определим функцию $P_*: 2^X \to L$ для каждого $D \subseteq X$ равенством:

$$P_*(D) = \sup\{P(A) \mid A \in \mathbf{A}, P(A) > P^*(A \setminus D)\}$$

(в этой записи предполагается, что $\sup \varnothing = 0$).

Лемма 1. (О продолжении меры возможности с условием).

Пусть D – подмножество X, $\delta \in L$. Тогда P можно продолжить до меры возможности $\bar{P}$ на 2^X такой, что $\bar{P}(D) = \delta$ тогда и только тогда, когда $\delta \le P^*(D)$ и выполняется хотя бы одно следующих условий:

1) если A, $(B_t)_{t \in T}$ – множества из класса $\mathbf{A}$ такие, что $A \subseteq D \cup \bigcup_{t \in T} B_t$, то $P(A) \le \delta \vee \sup\limits_{t \in T} P(B_t)$;

2) $\forall A \in \mathbf{A}\ P(A) > \delta \Rightarrow P^*(A \setminus D) = P(A)$;

3) $P_*(D) \le \delta$.

Доказательство. Докажем утверждение леммы для условия 1.

Необходимость. Предположим, что продолжение $\bar{P}$ существует. Выберем множества $A, B_t \in \mathbf{A}$, такие, что $A \subseteq D \cup \left(\bigcup\limits_{t \in T} B_t \right)$. Тогда

$$\bar{P}(A) \le \bar{P}(D) \vee \sup_{t \in T} \bar{P}(B_t) \le \delta \vee \sup_{t \in T} P(B_t).$$

Также, $\delta = \bar{P}(D) \le P^*(D)$.

Достаточность. Положим $\mathbf{A}^D = \mathbf{A} \cup \{D\}$. Определим функцию P на классе $\mathbf{A}^D$ равенствами $P(D) = \delta$ и $P(A) = P_0(A)$ при $A \in \mathbf{A}$. Пусть A^D и $(A_t^D)_{t \in T}$ – элементы $\mathbf{A}^D$. Докажем, что из включения $A^D \subseteq \bigcup\limits_{t \in T} A_t^D$ следует $P(A^D) \le \sup\limits_{t \in T} P(A_t^D)$.

Для этого рассмотрим 4 случая:

1) элементы A^D и $(A_t^D)_{t \in T}$ принадлежат $\mathbf{A}$. Тогда $A^D = \bigcup\limits_{t \in T} \left(A_t^D \cap A^D \right)$ и

$$P(A^D) = \sup_{t\in T} P(A_t^D \cap A^D) \le \sup_{t\in T} P(A_t^D).$$

2) $A^D = D$, а элементы $(A_t^D)_{t\in T}$ принадлежат $\mathbf{A}$. Тогда множества A_t^D образуют покрытие множества D, поэтому $P(A^D) = \delta \le P^*(D) \le \sup_{t\in \mathbf{A}} P(A_t^D)$.

3) $A^D \in \mathbf{A}$ и среди элементов $(A_t^D)_{t\in T}$ есть множество D. Пусть $T^0 \subseteq T$ – множество индексов t, таких, что $A_t^D \in \mathbf{A}$ (возможно пустое). Тогда по условию леммы:

$$P(A^D) = P_0(A^D) \le \delta \vee \sup_{t\in T^0} P_0(A_t^D) = \sup_{t\in T} P(A_t^D).$$

4) $A^D = D$ и среди элементов $(A_t^D)_{t\in T}$ есть множество D. Тогда $P(A^D) \le \sup_{t\in T} P(A_t^D)$.

Таким образом, выполняются условия теоремы о продолжении меры возможности для функции P на классе $\mathbf{A}^D$, поэтому существует продолжение P до меры возможности $\bar{P}$ на 2^X. Мера возможности $\bar{P}$ является продолжением P_0 и удовлетворяет условию $\bar{P}(D) = \delta$.

Докажем, что из условия 1 следует условие 2.

Пусть $A \in \mathbf{A}$ и $P(A) > \delta$. Пусть $(B_t)_{t\in T}$ – произвольное покрытие множества $A \setminus D$ элементами класса $\mathbf{A}$. Тогда $A \setminus D \subseteq \bigcup_{t\in T} B_t$, $A \subseteq D \cup \bigcup_{t\in T} B_t$. И по условию 1: $P(A) \le \delta \vee \sup_{t\in T} P(B_t)$. Учитывая, что $P(A) > \delta$, получаем $P(A) \le \sup_{t\in T} P(B_t)$.

Таким образом, $P(A) \le P^*(A \setminus D)$, откуда $P^*(A \setminus D) = P(A)$.

Докажем, что из условия 2 следует условие 3.

Из условия 2 следует, что для любого множества $A \in \mathbf{A}$, такого, что $P(A) > P^*(A \setminus D)$ выполняется неравенство $P(A) \le \delta$. Тогда $P_*(D) = \sup\{P(A) \mid A \in \mathbf{A}, P(A) > P^*(A \setminus D)\} \le \delta$.

Докажем, что из условия 3 следует условие 1.

Пусть $P_*(D) \le \delta$ и A, $(B_t)_{t\in T}$ – множества из класса $\mathbf{A}$, такие, что $A \subseteq D \cup \bigcup_{t\in T} B_t$.

Тогда $A \setminus D \subseteq \bigcup_{t \in T} B_t$, т.е. множества $(B_t)_{t \in T}$ образуют покрытие множества $A \setminus D$, поэтому $P^*(A \setminus D) \le \sup_{t \in T} P(B_t)$. Если $P(A) \le P^*(A \setminus D)$, то $P^*(A \setminus D) \le \sup_{t \in T} P(B_t)$.

Если же $P(A) > P^*(A \setminus D)$, то по условию 3, $P(A) \le \delta$. В обоих случаях выполняется неравенство:

$$P(A) \le \delta \vee \sup_{t \in T} P(B_t).$$

Лемма доказана.

Следствие. Если $A \in \mathbf{A}$, то $P_*(A) = P(A)$.

Доказательство. Поскольку P имеет продолжение до меры возможности на 2^X и любое такое продолжение $\bar{P}$ удовлетворяет условию $\bar{P}(A) = P(A)$, то $P_*(A) = P(A) = P^*(A)$.

Примечание. Функция P_* может не быть мерой возможности, как показывает следующий пример. Положим $X = \{0,1\}$, $\mathbf{A} = \{\varnothing, X\}$, $P(\varnothing) = 0$, $P(X) = 1$. Тогда $P^*(\{0\}) = P^*(\{1\}) = 1$ и $P_*(\{0\}) = P_*(\{1\}) = 0$, но $P_*(\{0,1\}) = 1$.

Определение 8. Распределение $Q \in FD_M$ называется вырожденным, если $Q(\{y\}) > 0$ не более чем для одного элемента $y \in M$.

Утверждение 3. Выполняются следующие свойства:

1) каждое распределение $Q \in FD_M$ имеет модель;

2) каждое распределение последовательности $Q_\omega \in FS_M^\omega$ имеет модель;

Доказательство.

1) Положим $X = \{(y,p) \mid y \in M, p = Q(\{y\})\}$, $P(A) = \sup_{(y,p) \in A} p$, $\xi((y,p)) = y$.

Тогда $P_\xi(Y) = P\{(y,p) \mid \xi((y,p)) \in Y\} = \sup\{Q\{y\} \mid y \in Y\} = Q(Y)$, $Y \subseteq M$.

2) Доказательство аналогично пункту 1.

Утверждение доказано.

Лемма 2. Пусть Q_+ – система конечномерных распределений последовательности. Тогда каждая ее модель является моделью сходимости с необходимостью 1 тогда и только тогда, когда для каждой расходящейся последовательности (y_n) выполняется $\lim\limits_{n\to\infty} Q_+(\{(y_1,..,y_n)\})=0$.

Доказательство. Необходимость. Предположим, что существует расходящаяся последовательность $y^0=(y_1^0,y_2^0,..)\in M^\omega$, такая, что $\delta=\lim\limits_{n\to\infty} Q_+(\{(y_1^0,..,y_n^0)\})>0$.

Пусть Q^* – внешняя мера возможности, соответствующая мере возможности $Inf(Q_+)$. Тогда $Q^*(\{y_{(\cdot)}\})=\inf\limits_{n>0} Inf(Q_+)((y_1,..,y_n)*M^\omega)=\lim\limits_{n\to\infty} Q_+\{(y_1,..,y_n)\}$ для произвольного $y_{(\cdot)}\in M^\omega$. Отсюда $Q^*(\{y^0\})=\delta>0$, и поскольку Q^* имеет модель, которая является моделью Q_+, то Q_+ имеет модель которая, не является моделью сходимости с необходимостью 1, что противоречит предположению и завершает доказательство необходимости.

Достаточность. Для каждой расходящейся последовательности (y_n) выполняется равенство $\lim\limits_{n\to\infty} Q_+(\{(y_1,..,y_n)\})=0$, и поэтому $Q^*(\{(y_1,y_2,..)\})=0$. Поскольку для произвольного продолжения Q_ω меры возможности $Inf(Q_+)$ на булеан множества M^ω выполняется неравенство $Q_\omega\{(y_1,y_2..)\}\le Q^*\{(y_1,y_2..)\}$, то $Q_\omega\{(y_1,y_2..)\}=0$. Следовательно, произвольная модель системы конечномерных распределений Q_+ является моделью сходимости с необходимостью 1.

Лемма доказана.

Следствие. Если пространство M полно и ограничено, то условие леммы можно заменить таким: $Q_+\{(y_1,..,y_N)\}\sup\limits_{n,m>N} d(y_n,y_m)\to 0$ при $N\to\infty$ для произвольной последовательности (y_n).

Доказательство. Пусть для каждой расходящейся последовательности (y_n) выполняется равенство $\lim\limits_{n\to\infty} Q_+(\{(y_1,..,y_n)\})=0$. Тогда

$$\lim_{N\to\infty} Q_+\{(y_1,..,y_N)\}\sup_{n,m>N} d(y_n,y_m)=\lim_{N\to\infty} Q_+\{(y_1,..,y_N)\}\lim_{N\to\infty}\sup_{n,m>N} d(y_n,y_m),$$

поскольку пространство M ограничено.

Если последовательность y_n сходится, то из полноты пространства M получаем равенство $\lim\limits_{N\to\infty}\sup\limits_{n,m>N} d(y_n, y_m) = 0$.

Если последовательность y_n расходится, то $\lim\limits_{N\to\infty} Q_+\{(y_1,..,y_N)\} = 0$. В обоих случаях

$$Q_+\{(y_1,..,y_N)\}\sup_{n,m>N} d(y_n, y_m) \to 0.$$

Наоборот, если $Q_+\{(y_1,..,y_N)\}\sup\limits_{n,m>N} d(y_n, y_m) \to 0$ для каждой последовательности (y_n), то для каждой расходящейся последовательности (y_n) выполняется неравенство $\lim\limits_{N\to\infty}\sup\limits_{n,m>N} d(y_n, y_m) > 0$, откуда $Q_+\{(y_1,..,y_N)\} \to 0$.

Следствие доказано.

Как показывает следующий пример, условие леммы 2 не эквивалентно условию существования модели сходимости по возможности (т.е. наличие модели сходимости по возможности является лишь достаточным условием для того, чтобы, каждая модель сходимости являлась моделью сходимости с необходимостью 1).

Пример 1. Система конечномерных распределений последовательности, которая не имеет модели сходимости по возможности, и при этом все ее модели являются моделями сходимости с необходимостью 1.

Положим $M = \{0,1\}$ с дискретной метрикой и

$$Q_+(\{(y_1,..,y_n)\}) = \begin{cases} 1, & y_1 + .. + y_n \le 1, \\ 0, & y_1 + .. + y_n > 1. \end{cases}$$

Тогда в каждой расходящейся последовательности (y_n) имеется более 1 единицы и, следовательно, $\lim\limits_{n\to\infty} Q_+(\{(y_1,..,y_n)\}) = 0$. По лемме 2, каждая модель Q_+ является моделью сходимости с необходимостью 1.

Пусть последовательность нечетких персептивных элементов ξ_n имеет распределение Q_+. Тогда для каждого $n \geq 1$, $P\{\sup_p d(\xi_{n+p}, \xi_n) \geq 1\} \geq Q_+(\{0^n 1\}) = 1$, поэтому $P\{\sup_p d(\xi_{n+p}, \xi_n) \geq 1\} \to 1$ при $n \to \infty$ и, следовательно, последовательность ξ_n не является фундаментальной по возможности. Поскольку пространство M полное, то ξ_n не является сходящейся по возможности. Следовательно, Q_+ не имеет модели сходимости по возможности.

Лемма 3. Пусть Q_+ – система конечномерных распределений последовательности. Она имеет модель сходимости с необходимостью 1 тогда и только тогда, когда для каждого $\delta > 0$, $k \in N$ и элементов $y_1, .., y_k \in M$ существует сходящаяся последовательность $y_{k+1}, y_{k+2}, ...$ такая, что

$$\lim_{n \to \infty} Q_+\{(y_1, .., y_n)\} + \delta > Q_+\{(y_1, .., y_k)\}.$$

Доказательство. Применим лемму 1 к случаю, когда $X = M^\omega$, $\mathbf{A} = BT_M^0$, D – множество расходящихся последовательностей. Меру возможности P положим равной $Inf(Q_+)$. Тогда $P^*\{(y_1, y_2..)\} = \lim_{n \to \infty} Q_+\{(y_1, .., y_n)\}$, и условием существования модели сходимости с необходимостью 1 для Q_+ является условие $\forall A \in \mathbf{A}$ $P^*(A \setminus D) = P(A)$, то есть каждого $k \in N$ и элементов $y_1, .., y_k \in M$, $\sup\{\lim_{n \to \infty}\{(y_1, .., y_n)\} \mid (y_{k+1}, y_{k+2}, ..)$ сходящаяся последовательность$\} = Q_+\{(y_1, .., y_k)\}$. Отсюда получаем условие леммы.

Лемма доказана.

Лемма 4. Пусть Q_+ – система конечномерных распределений последовательности. Тогда она имеет модель расходимости с необходимостью 1 тогда и только тогда, когда для каждого $\delta > 0$, $k \in N$ и элементов $y_1, .., y_k \in M$ существует расходящаяся последовательность $y_{k+1}, y_{k+2}, ...$, такая, что

$$\lim_{n \to \infty} Q_+\{(y_1, .., y_n)\} + \delta > Q_+\{(y_1, .., y_k)\}.$$

Доказательство аналогично доказательству предыдущей леммы.

Следующий пример показывает, что условия лемм 3 и 4 не являются взаимно исключающими.

Пример 2. Система конечномерных распределений последовательности, которая имеет модель сходимости с необходимостью 1 и модель расходимости с необходимостью 1.

Определим систему Q_+ таким образом, что $Q_+(\{(y_1,\ldots,y_n)\})=1$ для каждого $n\geq 1$ и $y_1,..,y_n \in M$. Тогда Q_+ является системой конечномерных распределений последовательности. Поскольку для произвольной последовательности y_n и индекса k выполняется

$$\lim_{n\to\infty} Q_+(\{(y_1,\ldots,y_n)\}) = Q_+\{(y_1,\ldots,y_k)\},$$

то по леммам 3 и 4, Q_+ имеет как модель сходимости с необходимостью 1, так и модель расходимости с необходимостью 1.

Применим полученные выше результаты для того, чтобы определить, имеет ли место теоретико-возможностный аналог закона больших чисел.

Определение 9. Две последовательности нечётких персептивных элементов $\xi_n : X \to R$ и $\xi'_n : X' \to R$, заданные на пространствах возможностей $(X,2^X,P)$ и $(X',2^{X'},P')$ называются эквивалентными по конечномерным распределениям, если для любых $n\geq 1$, $y_1,\ldots,y_n \in R$ выполняется

$$P\{\xi_1 = y_1,\ldots,\xi_n = y_n\} = P'\{\xi'_1 = y_1,\ldots,\xi'_n = y_n\}$$

Теорема 1. Пусть ξ_n – последовательность независимых одинаково распределенных нечетких персептивных элементов на пространстве возможностей $(X,2^X,P)$. Тогда:

1) существует пространство возможностей $(X',2^{X'},P')$ и последовательность нечетких персептивных элементов ξ'_n на нем, эквивалентная по конечномерным распределениям последовательности ξ_n такая, что последовательность $\frac{1}{n}\sum_{i=1}^{n}\xi'_i$ сходится с необходимостью 1;

2) если распределение f_ξ персептивного элемента ξ_1 не вырождено, то существует пространство возможностей $(X'',2^{X''},P'')$ и последовательность нечетких персептивных элементов ξ''_n на нем, эквивалентная по конечномерным распределениям последовательности ξ_n такая, что последовательность $\frac{1}{n}\sum_{i=1}^{n}\xi'_i$ расходится с положительной необходимостью.

Доказательство.

1) Положим Q_+ – система конечномерных распределений последовательности $\eta_n = \frac{1}{n}\sum_{i=1}^{n}\xi_i$.

a) Воспользуемся леммой 3. Пусть выбрано произвольное $\delta > 0$, $k \in N$ и элементы $z_1,\ldots,z_k \in M$. Поскольку $\sup_y f_\xi(y) = 1$, то существует y^* , для которого $f_\xi(y^*) > 1-\delta$. Положим $z_{j+1} = \frac{y^* + jz_j}{j+1}$ при $j \geq k$.

b) При $n > k$, $Q_+\{(z_1,\ldots,z_n)\} = P\{\xi_i = y_i, i = 1,\ldots,n\}$, где $z_i = \frac{1}{i}\sum_{j\leq i} y_j \;\forall i = 1,\ldots,n$.

Поэтому $y_i = iz_i - (i-1)z_{i-1}$. Учитывая независимость ξ_n и определение элементов $z_{j+1}, j \geq k$, получаем:

$$Q_+\{(z_1,\ldots,z_n)\} = f_\xi(z_1) \wedge f_\xi(2z_2 - z_1) \wedge \ldots \wedge f_\xi(nz_n - z_{n-1}) =$$

$$= Q_+\{(z_1,\ldots,z_k)\} \wedge f_\xi(y^*) \wedge \ldots \wedge f_\xi(y^*) .$$

Из последнего равенства следует, что

$$\lim_{n\to\infty} Q_+\{(z_1,\ldots,z_n)\} \geq Q_+\{(z_1,\ldots,z_k)\} \wedge (1-\delta) , \text{ и}$$

$$\lim_{n\to\infty} Q_+\{(z_1,\ldots,z_n)\} + \delta \geq Q_+\{(z_1,\ldots,z_k)\} .$$

Кроме того, $z_{j+1} - y^* = \frac{y^* + jz_j}{j+1} - y^* = \frac{j(z_j - y^*)}{j+1}$ при $j \geq k$, поэтому

$$\left|z_n - y^*\right| = \frac{n-1}{n}\frac{n-2}{n-1}\ldots\frac{k}{k+1}\left|z_k - y^*\right| = \frac{k}{n}\left|z_k - y^*\right| \to 0 \text{ при } n \to \infty .$$

Из чего следует сходимость последовательности z_n. Поскольку $\delta > 0$ и $z_1,..,z_k \in M$ были выбраны произвольно, то по лемме 3, Q_+ имеет модель $((X',P'),\eta'_n)$ сходимости с необходимостью 1.

Положим $\xi'_n = n\eta'_n - (n-1)\eta'_{n-1}, n \geq 1$. Для произвольных y_i выполняется равенство:

$$P'\{\xi'_i = y_i, i = 1,\ldots,n\} = P'\{\eta'_i = z_i, i = 1,\ldots,n\} = P\{\eta_i = z_i, i = 1,\ldots,n\} = P\{\xi_i = y_i, i = 1,\ldots,n\},$$

где $z_i = \sum_{j=1}^{i} y_j, i = 1,\ldots,n$.

Поэтому последовательность ξ'_n эквивалентна по конечномерным распределениям последовательности ξ_n.

2) Поскольку распределение ξ_1 не вырождено, то существуют вещественные числа a, b, $a \neq b$, такие, что $f_{\xi_1}(a) > 0$ и $f_{\xi_1}(b) > 0$. Построим последовательность $y = \otimes_{k\geq 1}(a^{2^k} * b^{2^k}) \in R^{\omega}$, где $\otimes$ и $*$ обозначают конкатенацию элементов для образования слова или ω-слова. Положим $z_n = \frac{1}{n}\sum_{i=1}^{n} y_i$. Тогда

$$Q_+(\{(z_1,\ldots,z_n)\}) \geq P\{\xi_i = y_i, i = 1,\ldots,n\} = \min_{i=1,..,n} f_{\xi}(y_i) = f_{\xi}(a) \wedge f_{\xi}(b),$$

и соответственно, $\lim_{n\to\infty} Q_+(\{(z_1,\ldots,z_n)\}) \geq f_{\xi}(a) \wedge f_{\xi}(b) > 0$.

Положим $n = 2(1 + 2 + 2^2 + \ldots + 2^s)$, $p = 2^{s+1}$. Тогда

$$\left|z_{n+p} - z_n\right| = \frac{1}{n+p}\left|\sum_{i=1}^{n+p} y_i - \frac{n+p}{n}\sum_{i=1}^{n} y_i\right| = \frac{1}{n+p}\left|\sum_{i=n+1}^{n+p} y_i - \frac{p}{n}\sum_{i=1}^{n} y_i\right| =$$

$$= \frac{1}{n+p}\left|pa - \frac{p}{n}\frac{n}{2}(a+b)\right| = \frac{p}{n+p}\frac{|a-b|}{2} = \frac{2^{s+1}}{2(2^{s+1}-1)+2^{s+1}}\frac{|a-b|}{2} = \frac{1}{3-1/2^s}\frac{|a-b|}{2}.$$

Поэтому для каждого n, $\sup\limits_{p>0}\left|z_{n+p}-z_n\right|\geq\dfrac{|a-b|}{6}$ и, следовательно, последовательность z_n расходится. Тогда по лемме 2, не каждая модель Q_+ является моделью сходимости с необходимостью 1 и, следовательно, Q_+ имеет модель $((X'',P''),\eta_n'')$ расходимости с положительной необходимостью. Аналогично пункту 1, делаем вывод, что последовательность $\xi_n''=n\eta_n''-(n-1)\eta_{n-1}'', n\geq 1$ эквивалентна по конечномерным распределениями последовательности ξ_n.

Теорема доказана.

Теорема 2. Пусть $\xi_n : X\to\mathbf{R}$ – последовательность независимых одинаково распределенных нечетких персептивных элементов на одном пространстве возможностей. Тогда последовательность $\eta_n=\dfrac{1}{n}\sum\limits_{i=1}^{n}\xi_i$ сходится по возможности тогда и только тогда, когда распределение ξ_n вырождено.

Доказательство. Необходимость. Предположим, что η_n сходится с необходимостью 1, но распределение ξ_n не вырождено. Тогда по критерию типа Коши для сходимости с необходимостью 1, выполняется соотношение

$$\forall c>0:\lim_{n\to\infty}P\left(\sup_{m>n}\left|\eta_m-\eta_n\right|>c\right)=0.$$

Пусть $f(y)$ – функция распределения нечёткого персептивного элемента ξ_n. Поскольку распределение не вырождено, то выберем пару точек $y_1\neq y_2$, для которых $f(y_i)>0, i=1,2$. Тогда при $m=2n$, для некоторого $\varepsilon>0$ выполняется неравенство:

$$P\left\{\left|\eta_m-\eta_n\right|\geq\left|y_1-y_2\right|/2\right\}\geq P\{\eta_n=y_1,\eta_m=(y_1+y_2)/2\}\geq$$

$$\geq P\{\xi_1=..=\xi_n=y_1,\xi_{n+1}=..=\xi_{2n}=y_2\}=\min\{f(y_1),f(y_2)\}>\varepsilon.$$

Положив $c = |y_1 - y_2| / 4$, получаем, что для каждого $n \geq 1$ существует $x_n \in X_\varepsilon$, для которого $\sup_{m>n} |\eta_m(x_n) - \eta_n(x_n)| > c$. Следовательно, $\underline{\lim}_{n\to\infty} P\left(\sup_{m>n} |\eta_m - \eta_n| > c \right) \geq \varepsilon$ – противоречие. Таким образом, распределение ξ_n вырождено.

Достаточность. Если $P\{\xi_n \in M\} = 0$, то утверждение очевидно. Пусть распределение ξ_n вырождено, $P_{\xi_n}(\{y\}) = 0$ и $\forall y \neq y_0, P_{\xi_n}(\{y_0\}) > 0$. Тогда нечёткие персептивные элементы ξ_n и η_n равны с необходимостью 1 константе y_0, поэтому $P\left(\sup_{m>n} |\eta_m - \eta_n| > c \right) = 0$ при $c > 0$.

Теорема доказана.

Теоремы 1 и 2 показывают, что для сходимости по возможности и для сходимости с необходимостью 1, теоретико-возможностный аналог закона больших чисел не выполняется.

Заключение

В статье получены критерии существования модели сходимости и расходимости с необходимостью 1 для систем конечномерных распределений последовательностей нечётких персептивных элементов (леммы 2-4). Кроме того, доказаны теоремы о не выполнении теоретико-возможностного аналога закона больших чисел для сходимости по возможности и сходимости с необходимостью 1.

Литература

[Boyel,1995] L.Boyel, G. de Cooman, E. E. Kerre. On the extension of P-consintent mappings // Proc. FAPT '95, Gent, 1995. – pp. 88 – 98

[Zadeh, 1978] L.A. Zadeh. Fuzzy sets as a basis for a theory of possibility // Fuzzy Sets and Systems, 1978. Vol. 1, pp. 3-28.

[Wang, 1992] Z.Wang, G. J. Klir. Fuzzy Measure Theory. Plenum Press, New York, 1992.

[Бычков, 2007a] А.С. Бычков, К.С.Колесников. Построение (PN)-модели теории возможностей // Вестник Киевского университета, Серия: физико-математические науки, 2007. №1, с. 134-138.

[Бычков, 2007b] А.С. Бычков. Об одном развитии теории возможностей // Кибернетика и системный анализ, 2007. №5, с. 67-72.

[Дюбуа, 1990] Д.Дюбуа, А.Прад. Теория возможностей. Приложения к представлению знаний в информатике. — М.: Радио и связь, 1990. — 288 С.

[Пытьев, 2000] Ю.Пытьев. Возможность. Элементы теории и применение. УРСС,2000. —192 С.

[Пытьев, 2004] Ю.Пытьев. Неопределённые нечёткие модели и их применения // Интеллектуальные системы, 2004. Т. 8, вып. 1-4, с.147-310.

Authors' Information

Алексей Бычков *– к.ф.-м.н., заведующий кафедрой программирования и компьютерной техники факультета информационных технологий Киевского национального университета имени Тараса Шевченка; ул. Ломоносова 81А, 03022, Киев, Украина; e-mail: bos.knu@gmail.com*

Основная область научных интересов: исследование гибридных автоматов как моделей непрерывно-дискретных процессов; построение согласованной теории возможностей, нечетких персептивных величин и процессов; математические основы моделирования нечетких сложных систем; применение математических методов в биологии, медицине и экономике

Евгений Иванов *– к.ф.-м.н., ассистент кафедры программирования и компьютерной техники факультета информационных технологий Киевского национального университета имени Тараса Шевченка; ул. Ломоносова 81А, 03022, Киев, Украина; e-mail: ivanov.eugen@gmail.com*

Основная область научных интересов: семантика языков программирования; формальные методы; математическая теория систем; гибридные (дискретно-неперервыеі) системы

Ольга Супрун *– к.ф.-м.н., доцент кафедры программирования и компьютерной техники факультета информационных технологий Киевского национального университета имени Тараса Шевченка; ул. Ломоносова 81А, 03022, Киев, Украина; e-mail: o.n.suprun@gmail.com*

Основная область научных интересов: математическое моделирование и вычислительные методы; нечеткие величины и процессы; гибридные модели непрерывно-дискретных процессов

About convergence of fuzzy perceptive elements sequences, defined on different opportunities spaces

Alexei Bychkov, Eugene Ivanov, Olha Suprun

Abstract: Criteria for the existence of a model of convergence and divergence with the need 1 for systems of finite-element sequences of fuzzy perceptive elements are obtained in this paper. Theorems about failing of the theoretic-possibility analogue of the law of large numbers for converge with the opportunities and the converge with need 1 are proven.

Keywords: *theory of possibility, fuzzy logic, convergence of fuzzy perceptive elements.*

AN APPROACH TO MULTIFACETED BUSINESS PROCESS MODELING WITH MODEL TRANSFORMATION TOOLS

Roman Nesterov, Lyudmila Lyadova

Abstract: *The approach to models generation automation and implementation of multifaceted business process modeling on the basis of graphical model transformation is described. To create graphical models of diverse notations (diagrams in notations of visual modeling languages) one can exploit visual modeling software tools and language workbenches, DSM platforms. Domain specific modeling tools allow simplifying model design process, to involve domain experts (they are not masters of information technologies and have not programming skills) to formal model development. Newly-created models can be converted into simulation models or specific analytical models with the model transformation tools. Therefore, at new task solving process with modeling tools modelers have not to duplicate model development with new tools in new language notation. Model designers can use most suitable tools and most expressive languages for models development in their domain to solve their tasks. Obtained models after transformation can be examined with means of specific simulation modeling systems including, for instance, AnyLogic, or with mathematical software packages such as Mathcad, Maple or Mathematica. The visual business process modeling notation choice is substantiated. Mathematical model named DFD-graph is used as mathematical basis of model generation tools. The normalization rules form the backbone to the DFD business process model normalization. This algorithm is the basis of automating model generation software implementation.*

Keywords: *business process modeling, visual modeling languages, business process analysis, mathematical modeling, model development automation, model transformations, model reusing.*

ACM Classification Keywords: *D.2 SOFTWARE ENGINEERING: D.2.2 Design Tools and Techniques – Computer-aided software engineering (CASE), Programmer workbench; D.2.13 Reusable Software – Domain engineering, Reuse models. I.6 SIMULATION AND MODELING: I.6.2 Simulation Languages; I.6.3 Applications; I.6.4 Model Validation and Analysis; I.6.5 Model Development. G.4 MATHEMATICAL SOFTWARE: Algorithm design and analysis, User interfaces.*

Introduction

The efficient company management is impossible without employing informational and analytical systems being created and functioned following the business process models developed by analysts. The process of modeling itself has two tasks: model development (in terms of visual language notations or mathematical constructs) and model examination with tools meeting the needs identified by analysts and supporting the decision-making process. Analysts and domain experts need a great variety of analytical tools for the multifaceted exploration of business systems and processes via different models expressing various aspects to be studied to solve decision-making tasks. These models are based on the formal notations varied in form or in mathematical apparatus suitable for analytical tasks solving.

Nowadays there exists a lot of diverse visual (graphical) notations one can use to build visual business process models including, for example, IDEF0 notation used for identifying causal order of operations; DFD – for data flows representation; eEPC – for event-driven processes description et cetera. Different instruments and graphical model editors can be exploited for the development of visual business process models.

The modeling software choice often determines the modeling language or formal notation to be used while developing visual business process models. The same systems and processes can be described using various languages regarding the primary modeling objectives. Thus, analysts are forced to build models of object in question all over again while moving among modeling objectives achievement.

To conduct holistic model analysis by means of mathematical software toolkits (Matlab, Maple, Mathcad etc.) it is necessary to develop analytical models in terms of corresponding constructs and formats being utilized in those tools listed above. Moreover, analytical or numerical model development corresponding to high dimension systems to perform the analysis employing those mathematical software packages can prove to be non-trivial task: apart from model scope itself, another reason making the movement from visual business process model to its analytical representation can be the lack of data required for full-fledged model development. Notice, however, that having the high dimension visual business process model, described with means of chosen graphical notation, the movement to a corresponding analytical model representation being applicable for further examination can be automated.

There are partial solution to the problem of transformation of process and system models based on Petri net exploitations [Dorrer, 2011; Zhow, 2015], simulation modeling techniques and tools [Lantsev, 2013] and specific visual model transformation mechanisms [Poryazov, 2005; Poryazov, 2008; Poryazov, 2009].

This paper suggests to find a solution to the following tasks connected with model transformation aimed at further multifaceted business processes analysis execution:

1) to create the abstract graph model of business process relevant to the model proposed and developed by analyst in terms of the notation of the visual modeling language applied by domain expert to solve tasks of business process analysis;
2) to develop an algorithm transforming visual model, represented with chosen visual notation, into created on the first step abstract graph model;
3) to perform the transformation of the obtained abstract graph model into the classical Petri net;
4) to develop the queuing system model corresponding to the abstract graph model obtained on the first step.

These mathematical models (Petri net and queuing system) can give the comprehensive view of the business process in question, especially possible bottlenecks. In addition, their analysis can be automated and implemented with the help of mathematical software toolkits listed in this section.

An Approach Logic

Taking into account the existing transformation methods of visual business process models developed in the different notations, the general layout of abstract graph model of business process development is being proposed below. It is based on the model normalization and generic internal representation generation that can be transformed in its turn into different analytical representations according to the primary objectives.

The proposed approach includes following steps:

1. A visual business process models development by analysts in terms of chosen modeling language and notations relevant for analysis tasks.
2. An identification whether the visual model meets the requirements specified for the normalized visual model in a given notation.
3. The normalization process is to be implemented in case of any non-compliances identified at the previous step.
4. The generation of the business process model internal graph representation taking the normalized visual models as an input (model export).
5. The parse and loading the internal graph representation of the business process model into the environment of the chosen mathematical software tool.
6. An extension (refinement) of business process model definition (namely, determination of any control parameters of the model relevant to modeling goal) in the environment of the mathematical software

package exploited by analysts.

7. Automatically converting of the extended internal graph representation into an algebraic or differential equation system with means of tools offered by the chosen mathematical software package.
8. Finding the numerical or symbolic solution to the equation system obtained on the previous step.
9. A generation of a report on the business process analysis (the equation system solution output and analysis).

This proposed sequence of actions is supported by the implementation of algorithms aimed at the generation of analytical representation taking the visual business process model as an input.

The approach logic is shown in Fig. 1.

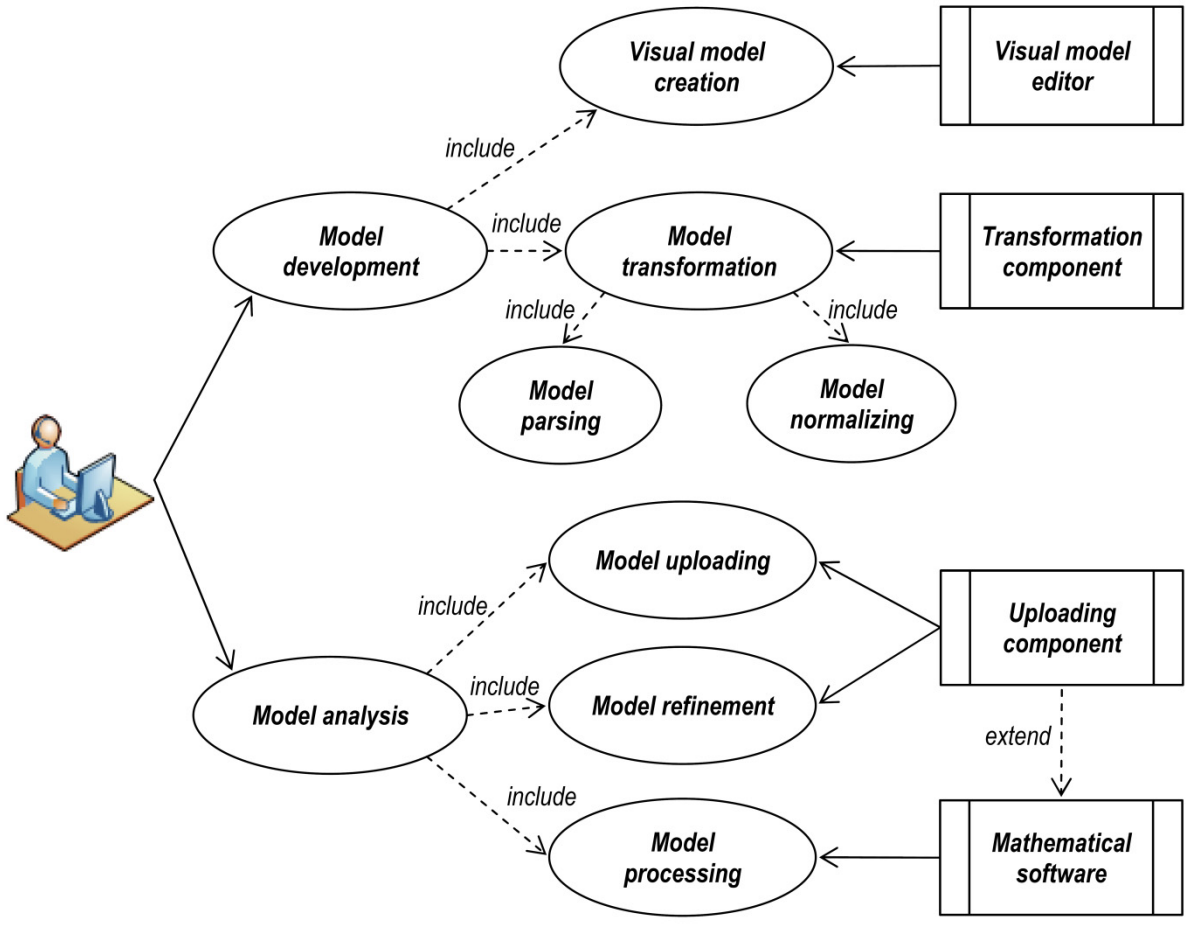


Figure 1. An approach to multifaceted business process modeling with model transformation tools and mathematical software

Furthermore, analysts can use DSM platforms to create domain specific languages (DSL) for the multifaceted business process modeling [Lyadova, 2014]. Domain specific modeling (DSM) tools allow to combine domain expert knowledge and model designer skills to develop and analyze models (Fig. 2). The language toolkits can become the basis for integration of various analytical tools (in particular, simulation systems). Maximal flexibility of modeling tools may be obtained with creating the multilevel models describing the researched systems and processes from various points of view and with different levels of details. For matching of various system descriptions it is necessary to develop the whole hierarchy of models (model, metamodel, meta-metamodel, etc.), where model is an abstract description of system characteristics that are important from the point of view of the modeling purpose, metamodel is a model of the language (DSL metamodel), which is used for models development, and meta-metamodel (metalanguage) is a language, on which metamodels are described in DSM platform.

Fig. 2 shows model development and analyzing with DSM platform MetaLanguage.

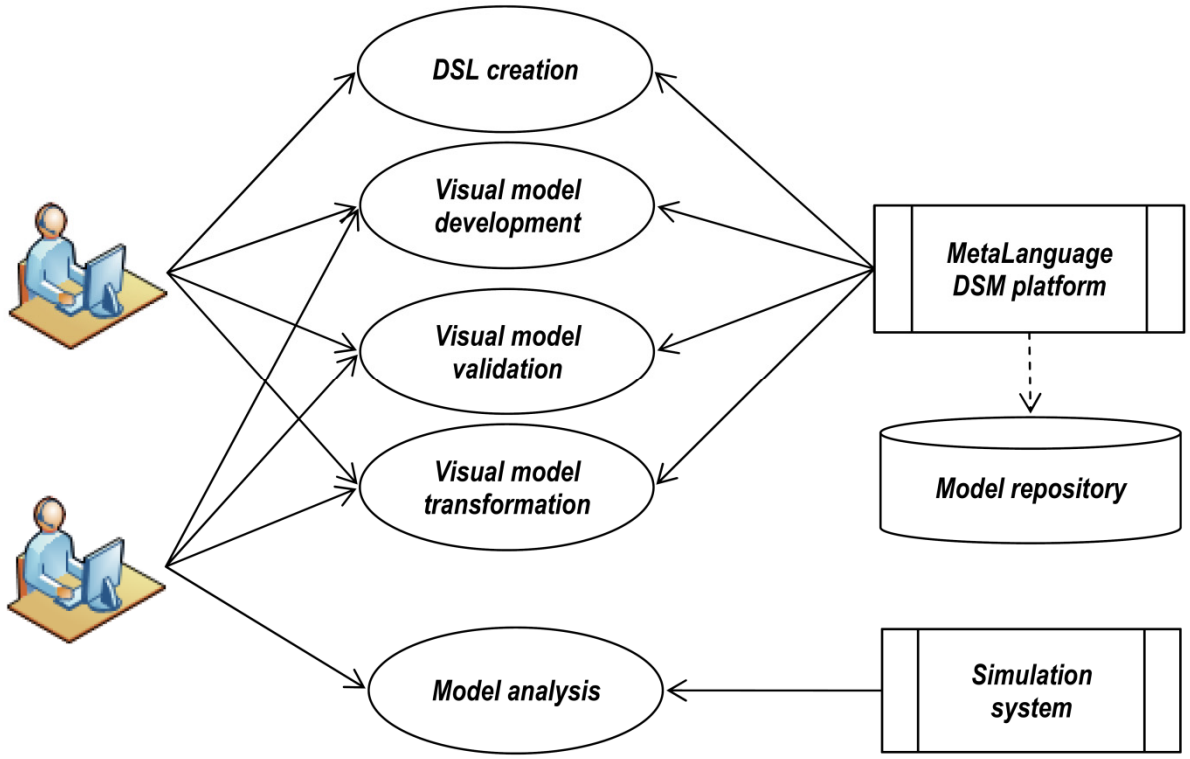


Figure 2. An approach to multifaceted business process modeling with model transformation tools and simulation systems

The Metalanguage system as an integration tool is presented in [Lyadova, 2014]. Transformations of business process models, described with MetaLanguage in various notations, into GPSS and AnyLogic models were realized.

The model editors and transformation tools allow to use various visual notations for business modeling. Therefore, analysts can choose more expressive and available modeling language to create models.

Business Process Modeling Notation Choice

To develop and pilot the proposed visual business process model transformation method to use mathematical software for model analysis it is necessary to choose the modeling language allowing an analyst to develop the model for business process of interest.

While choosing the graphical notation it is necessary to take into consideration the following:

1) the ability to yield the formal business process description;
2) the necessity to alter the developed model to implement the analysis with mathematical software toolkit instruments;
3) the ability to develop “end user friendly” business process model definition extension algorithm (business process control parameters specification).

Table 1 shows the results for comparison of different visual business process modeling notations against the listed criteria.

Table 1. The Comparison of Visual Business Process Modeling Notations

Notation	Possible indicators to examine	Mathematical apparatus
IDEF0	1) operation execution time; 2) actor time occupancy; 3) operation cost evaluation; 4) scenario probabilistic assessment	Algebraic equation systems

Notation	Possible indicators to examine	Mathematical apparatus
IDEF3	1) operation execution time; 2) operation cost evaluation; 3) scenario probabilistic assessment	Simulation modeling tools; Algebraic equation systems; Differential equation systems; Probabilistic evaluation models; Cols evaluation economical models
DFD	1) operation execution time; 2) operation cost evaluation; 3) a number of data access operation and time needed; 4) actor and external participant time occupancy; 5) data flow time characteristics; 6) the volume of data extracted and uploaded; 7) scenario probabilistic assessment	
BPMN	1) operation execution time; 2) operation cost evaluation; 3) actor time occupancy; 4) control flow time characteristics; 5) scenario probabilistic assessment	
eEPC	1) operation execution time; 2) operation cost evaluation; 3) actor time occupancy; 4) scenario probabilistic assessment	

According to the analysis results, DFD notation is chosen [Le Vie]. It is also chosen because it is not that “popular” notation exploited to model business processes and existing visual business process modeling tools offers too few tools to examine the models developed in terms of DFD notation. Analysts while building the models for business process mainly takes advantage of UML [Badreddin, 2010;

Bendraou, 2010; Hansen, 2012, *a*; Hansen, 2012, *b*; Mellor, 2002], IDEF3 or eEPC [Kim, 2001; Kim, 2003; Sterle, 2015]. These notations illustrate flow of operations, event chains and other related indicators. Nonetheless, an analyst can extract a lot of useful information that can be taken advantage of in the process of taking managerial decisions from extended DFD model defining it with control parameters.

The Main Business Process Indicators Identification

According to the DFD modeling capabilities main parameters values of which one can define or compute while processing the analytical representation of a visual business process models are listed below (the domain of interest is chosen regarding the usage of information resources while executing the business process defined with the help of DFD):

1) process is characterized by:
   a) execution time consisting of the time spent by all actors for executing this process operations;
   b) resource spend consisting of fixed and variable costs depending on the number of actors directly involved and time they spent;
   c) process run quantity generally defined by the input control parameters values and corresponding execution results;

2) flow is characterized by:
   a) speed/time of data movement (queries, requests, queries results et cetera) along the flow from one process to another or from a process to a data warehouse and vice versa;
   b) the amount of data moving across the flow of interest (only for flows connecting processes and warehouses);
   c) flow operation costs needed for data transmission organization including both fixed and variable ones;

3) data warehouse is characterized by:
   a) the number of extraction queries implemented by the processes;
   b) the number of update/insert queries implemented by the processes;
   c) the amount of data uploaded into the warehouses by the processes;
   d) the amount of data extracted from the warehouses by the process.

Apart from numerical indicators describing the specific business process DFD model parts, it is necessary to identify figures that can be computed using above listed control parameters as an input and that can help build the comprehensive picture of a business process in question, namely:

1) total execution time for one instance of a business process modeled with the help of DFD notation;
2) total amount of data uploaded and extracted from data warehouses according to queries done by the process parts of current instance of a business process in question;
3) total costs needed for the implementation process of one business process instance (including both fixed and variable costs).

It is necessary to mention that a part of all indicators cling to the currently studied business process is to be explicitly determined by the user. Thus, dependent parameters describing both single business process elements (processes, flows, data warehouses) and the business process on the whole is to be computed according to the mathematical and computing model described in the next section.

Mathematical Model of Business Process

The main formal definitions of business processes model obtained in terms of DFD visual modeling notation are listed below.

Definition 1. Graphical business process model M is a *DFD-model* if it can be represented as a directed marked graph as follows:

$$M = (P, D, F), \tag{1}$$

where $P = \{p_1, p_2, \ldots, p_n\}$ – a set of nodes representing *processes*, $D = \{d_1, d_2, \ldots, d_m\}$ – a set of nodes – *data warehouses*; and a set F represents *data flows* and being the union of two sets:

$$F = F_P \cup F_D, \tag{2}$$

where $F_P = \{f_{p1}, f_{p2}, \ldots, f_{pn}\}$ – represent flows between processes, $F_D = \{f_{d1}, f_{d2}, \ldots, f_{dn}\}$ – flows between processes and data warehouses.

Definition 2. Cross-process data flow in DFD-model M is a directed arc $f_{ij} = (p_i, p_j)$ from a subset F_P that meets: $p_i, p_j \in P$ and $p_i \neq p_j$.

So, the characteristic property for the elements of F_P is as follows:

$$F_P = \{(p_i, p_j) \mid p_i, p_j \in P, p_i \neq p_j\}. \quad (3)$$

Definition 3. "Process-Warehouse" data flow in DFD-model M is a directed arc $f_{ij} = (p_i, d_j)$ from the subset F_D meeting $p_i \in P$, $d_j \in D$.

Definition 4. "Warehouse-Process" data flow in DFD-model M is a directed arc $f_{ij} = (d_i, p_j)$ from the subset F_D meeting $d_i \in D$, $p_j \in P$.

The characteristic property for the elements of F_D is defined as follows:

$$F_D = \{ ((d_i, p_j) \mid d_i \in D, p_j \in P) \vee ((p_i, d_j) \mid p_i \in P, d_j \in D))\} \subset D \times P \cup P \times D. \quad (4)$$

Definition 5. Process p_i from a set P in DFD-model M is called *starting* if its in-degree is equal to 0, i.e. $d_{in}(p_i) = 0$. Starting process is defined by the means of subgraph $M_1 = (F_P, P)$.

Definition 6. Process p_i from a set P DFD-model M is called *terminating* if its out-degree is equal to 0, i.e. $d_{out}(p_i) = 0$. Starting process is also defined by the means of subgraph $M_1 = (F_P, P)$.

Definition 7. Business-process scenario defined via DFD-model M, is a path of the graph M connecting starting and terminating processes. Thus, scenario S_i ordered process subset $\{p_1, p_2, \ldots, p_k\}$ of the set of all processes P where each neighbor processes in S_i connected with cross-process data flow, i.e the set $\{(p_1, p_2), (p_2, p_3), \ldots, (p_{k-1}, p_k)\}$ is a subset to a set of all cross-process data flows F_P.

Fig. 3 shows the example of business process visual model, and Fig. 4 illustrates its corresponding abstract graph model M before additional marks for control parameters are introduced.

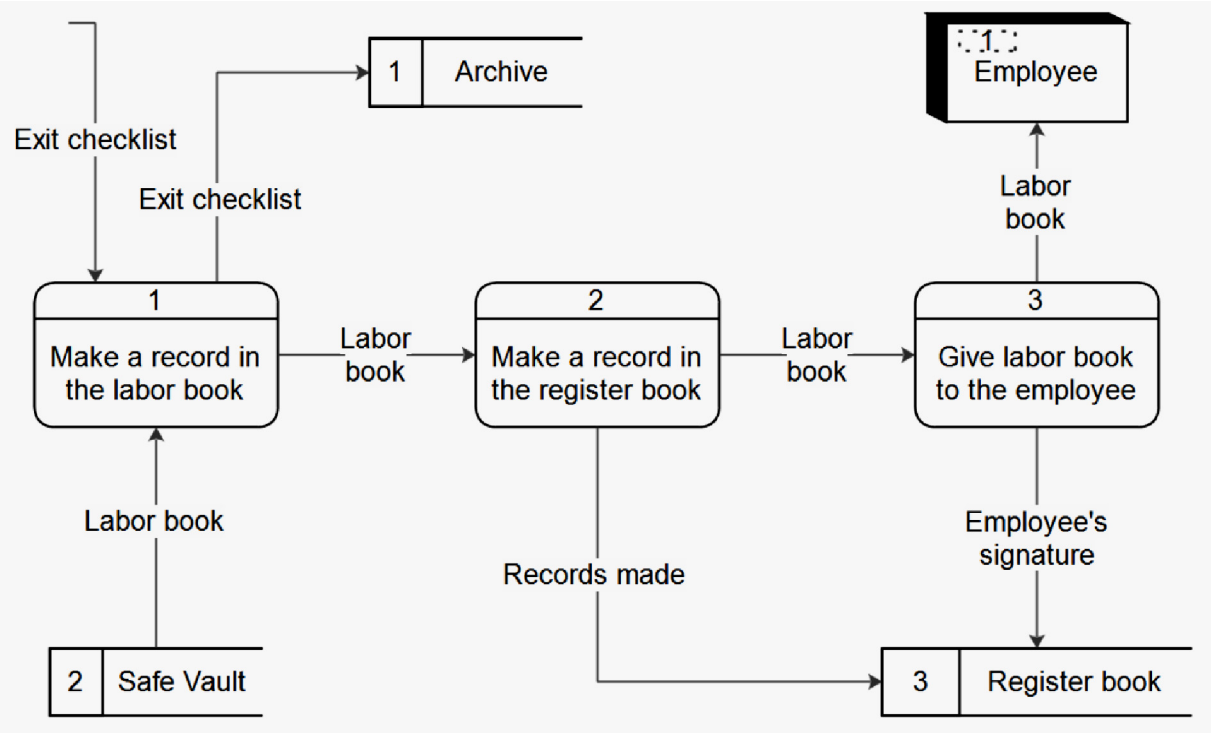


Figure 3. Visual DFD-model for a dismissal process

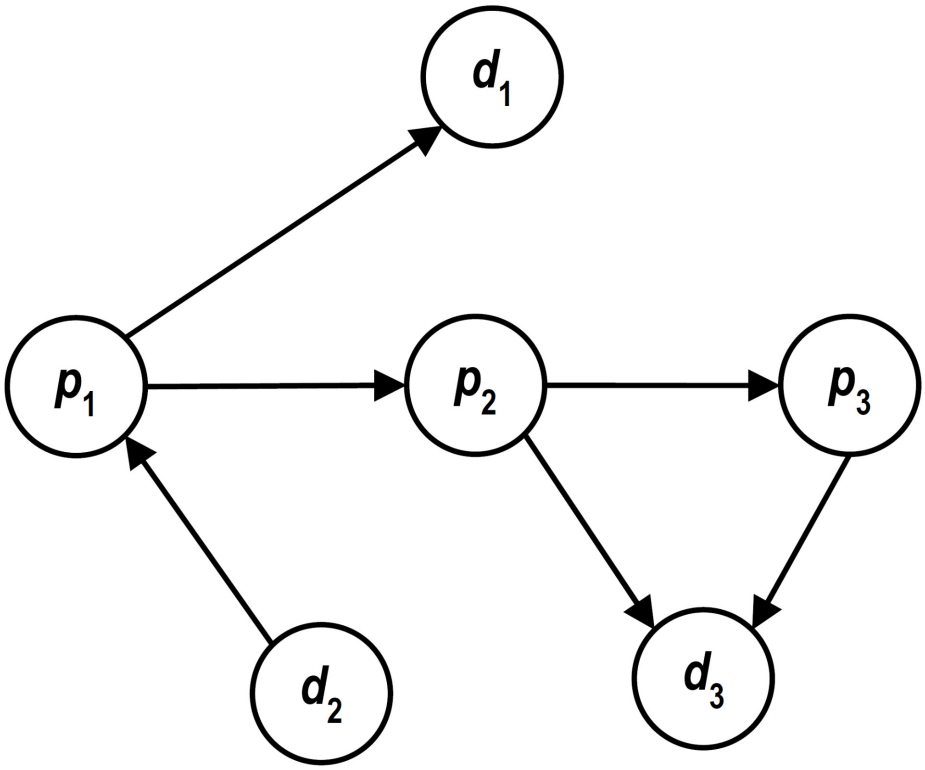


Figure 4. Dismissal process corresponding graph

Model Normalization Rules

As it was mentioned above, one of the main stages for the generation of business process analytical representation is normalization during that input model is checked to meet a set of predefined requirements. The set of normalization rules indicates the requirements for the visual DFD model of a business process providing the ability to develop the corresponding analytical representations pursuant to the formats listed in previous sections. Normalization rules is described in terms of definitions listed in the previous section.

Rule 1. DFD-model M can have only one starting process p_{start}. Otherwise, the model M is to be divided into several submodels having single starting processes.

Rule 2. DFD-model M can have at list one terminating process p_{finish}. Otherwise, it is necessary to introduce additional termination process node to the input graph M. The absence of a terminating process can be caused by graph cycles, that's why one need to introduce arc providing an exit from a cycles.

Rule 3. Each process of the DFD-model M is to be connected with another process p_j (at lest one), i.e.. Otherwise, the isolated process node not having connections with other process nodes is to be deleted.

Rule 4. Each process node p_i of the DFD-model M has to be connected with at least one data warehouse node d_j through the "Process-Warehouse" or "Warehouse-Process" data flows, i.e. $(p_i, d_j) \in F_D \vee (d_j, p_i) \in F_D$. Otherwise, the process node not having connections with warehouse nodes is to be deleted.

Rule 5. Each data warehouse node d_i of the DFD-model M is to have the connection with at least one process p_i across the "Process-Warehouse" or "Warehouse-Process" dataflows correspondingly. Otherwise, process node not having any connections with the process nodes is to be removed from the initial graph.

The normalization rules described above form the backbone to the DFD business process model normalization software implementation. Besides, the normalization algorithms is not described in this paper.

Figures Computing Technique

To facilitate the process of computing main process indicators and figures listed above, it is necessary to execute the additional business process abstract graph *M* transformation introducing arcs and nodes inscriptions (the input of this information is implemented by the user). Formal definitions to these parameters and definition extension algorithms will not be discussed in this paper.

To define the computing mechanism for business process indicators an initial abstract graph is to be divided into two subgraphs including a cross-process data flow subgraph $M_1 = (F_P, P)$ and a bipartite subgraph $M_2 = (F_D, P)$ indicating the interaction between processes and warehouses. Fig. 5 shows the example of graph division.

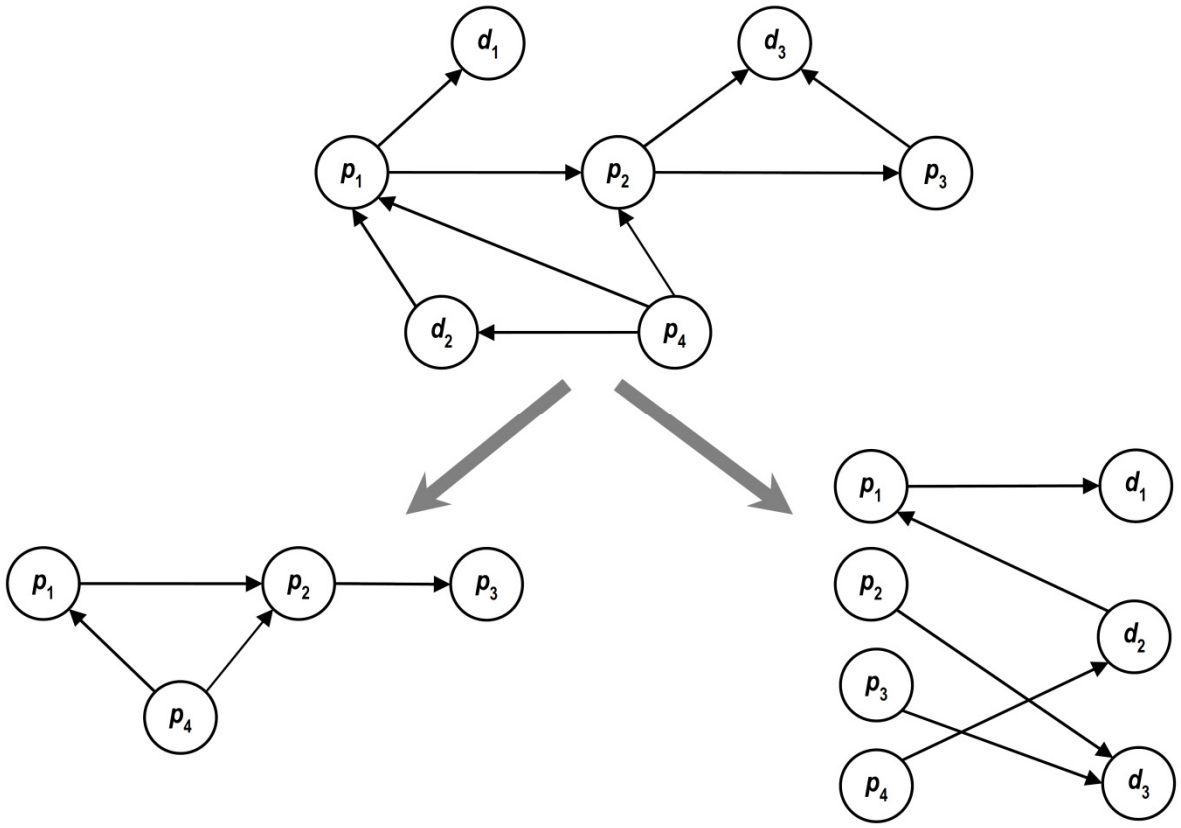


Figure 5. Initial business process abstract graph division

The cross-process data flow subgraph is used to analyze the business process implementation scenarios, and process-warehouse interaction subgraph is used to define and compute volume of information processed while business process instance operation.

As far as business process can be implemented through different scenarios, it is vital to inscribe each cross-process dataflow (p_i, p_j) with P_{ij} showing the probability of choosing this flow among the number of all available ones.

That is why, the total probability of business process implementation scenario S_i can be computed in the following way:

$$P(S_i) = \prod_{\substack{p_i, p_j \in S_i, \\ p_i \prec p_j}} (P_{ij}) \tag{5}$$

Formulas for each scenario characteristics computation are constructed (but these formulas aren't shown here):

- $T(S_i)$ – the execution time of the business process scenario S_i;
- $I(S_i)$ – the amount of information processed during implementation of the business process scenario S_i;
- $C(S_i)$ – the costs necessary for the execution of the business process scenario S_i.

By using the above mentioned extended graph and numerical indicators describing the single elements of the process, the following total business process (modeled with DFD-model M) figures computing algorithm were implemented:

- T_{total} – the total business process execution time;
- I_{total} – the total amount of information processed during the business process implementation;
- C_{total} – the total costs necessary for the execution of the business process in question.

The computing of DFD-model M above listed total figures, one need to get the values for the corresponding single business process element indicators.

Calculations can be executed with mathematical software.

DFD-model Transformation into a Classical Petri Net

To get more precise view on a possible problems that can occur while the business process instance implementation, one can exploit the apparatus offered by classical Petri net following the DFD-model transformation rules based on the correspondence of the DFD-model elements and Petri net constructs (see Table 2).

Table 2. Petri net and DFD-model correspondence

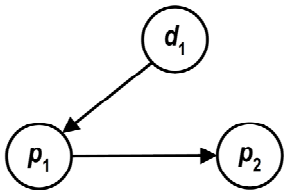
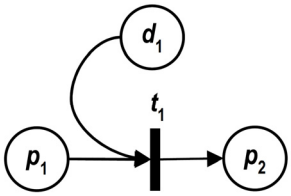
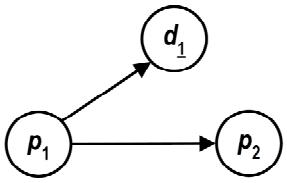
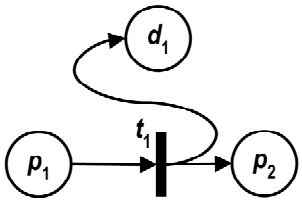
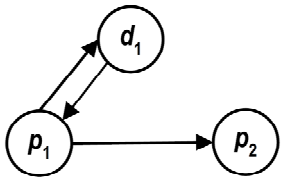
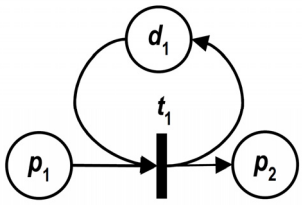
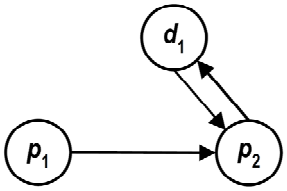
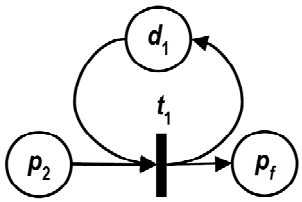
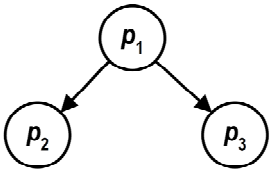
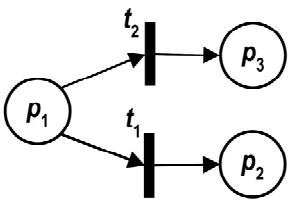
Element Meaning	DFD-model Elements	Petri Net Elements
Query data from a warehouse during process implementation		
Upload data to a warehouse during process implementation		
Simultaneous data query and upload while process implementation		
Data query and upload while terminating process implementation		
Alternative business process evolution		

Fig.7 shows the result of applying these transformation rules to a DFD-model abstract graph form a Fig.6.

The obtained Petri net will allow an analyst to perform additional analysis capabilities by means of Petri net specific algorithms: invariants, traps, deadlocks et cetera.

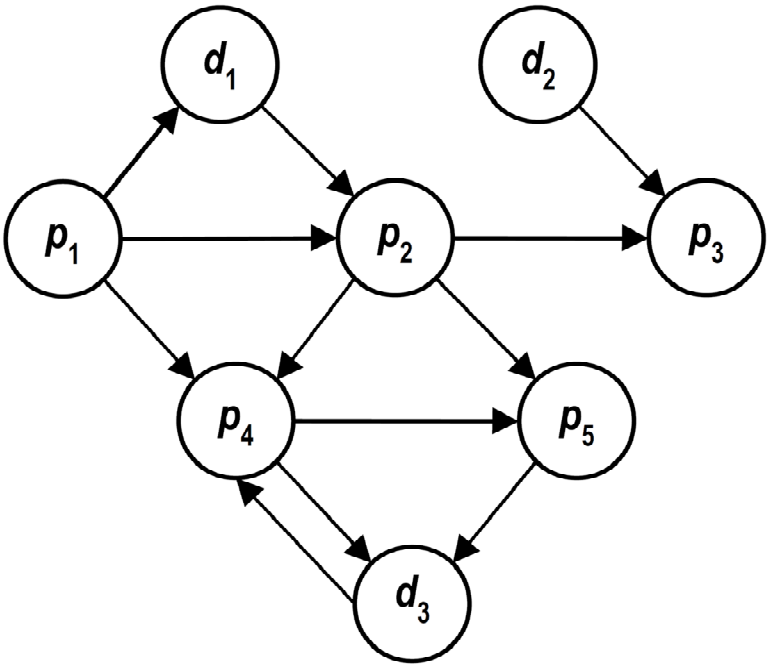


Figure 6. Abstract graph example of a DFD-model for a business process

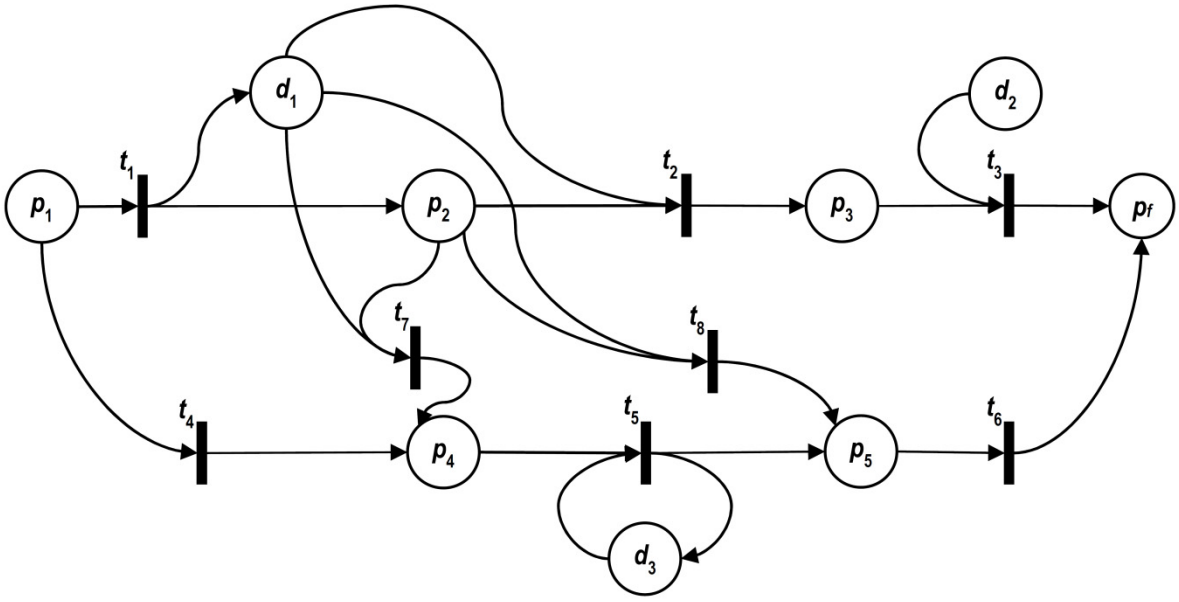


Figure 7. Corresponding Petri net

The generated Petri Net model can be uploaded to mathematical program to investigate its properties.

Queueing Model Application

Above described model do not resolve all the business process analysis problems. The other aspects concerning the business process modeled by using DFD notation can be examined with the help of queueing models, namely *process of destruction and multiplication*.

Within the framework of this model one can define the probabilities for a system being in one of the n possible states. The queuing system state S_i described by the occupancy of i queues from all n available ones.

The results obtained after the queueing model implementation can serve as a business process reengineering decision support. First of all, they can be applied to configure the queues work intensity, i.e. the probability of a system being in the idle state is to be decreased (when all queues are idle, or only the little part of them is occupied). Secondly, queueing modeling results can also help to redistribute the load among queues (more even possible system state probabilities distribution).

Software Implementation

Visual business process model transformation software research prototype has the following components implemented:

1) the visual business process model creation and its internal representation generation component;
2) the component for uploading and processing the internal representation into mathematical software tool Matlab.

The general scheme of the described above approach implementation is shown in Fig. 8.

The visual business process model creation and its internal representation generation component were implemented with the help of following instruments:

1) Microsoft Visual Studio 2012;
2) XML-file C# processing tools;
3) CASE-tool for automated model-driven development CASEBERRY (ICS company software);
4) ASP.NET MVC tools for model definition extension module development.

The component for uploading and processing the internal representation into mathematical software tool Matlab was implemented by means of following instruments:

1) Matlab Java integrated development environment;
2) Matlab graphical user interface development environment (GUIDE);
3) Java programming language XML-file processing tools.

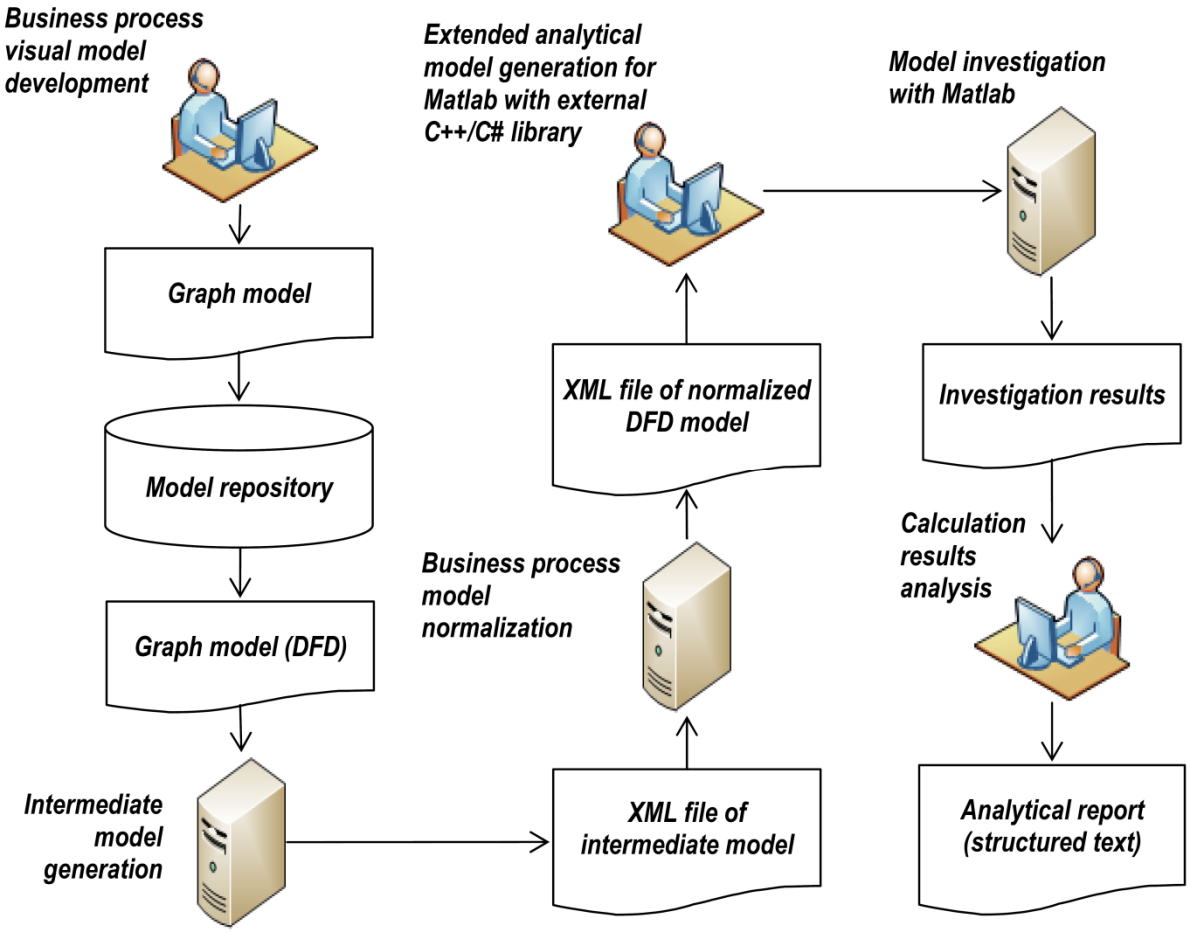


Figure 8. General implementation scheme

Conclusion

The research software prototype implemented shows the practical value for the proposed approach. The open system architecture allows to extend it with new components adding functionality including, for example, generation of other analytical representation types to study different business process aspects not presented in this paper so far.

It has to be mentioned that one can use DSM-platform MetaLanguage [Sukhov, 2013; Sukhov, 2014, *b*] to develop visual business process models of other types and notations and to transform and investigate diagrams.

Bibliography

[Badreddin, 2010] O.B. Badreddin. Umple: a Model-Oriented Programming Language. In: Proceedings of ACM/IEEE 32nd International Conference “Software Engineering”. 2010, vol. 2, pp. 337-338.

[Bendraou, 2010] R. Bendraou, J.-M. Jézéquel, M.P. Gervais, X. Blanc. A comparison of six UML-based languages for Foundation software process modeling. In: Software Engineering. 2010, vol. 36 (5), pp. 662-675.

[Hansen, 2012, *a*] H.H. Hansen, J. Ketema, B. Luttik, M.R. Mousavi, J. van de Pol. Towards model checking executable UML specifications in mCRL2. In: Innovations in Systems and Software Engineering. 2012, vol. 6 (1-2), pp. 83-90.

[Hansen, 2012, *b*] H.H. Hansen, J. Ketema, B. Luttik, M.R. Mousavi, J. van de Pol, O.M. dos Santos. Automated verification of executable UML models. In: Formal Methods for Components and Objects: Lecture Notes in Computer Science. Springer Berlin Heidelberg. Vol. 6957, 2012, pp. 225-250.

[Kim, 2003] C.-H. Kim, R.H. Westonb, A. Hodgsonb, K.-H. Lee. The complementary use of IDEF and UML modelling approaches. In: Computers in Industry. № 50, 2003, pp.35-56.

[Kim, 2001] C.-H. Kim, D.-S. Yim, R.H. Weston. An Integrated use of IDEF0, IDEF3 and Petri net methods in support of business process modelling. In: Proceedings of the Institution of Mechanical Engineers. Part E: Journal of Process Mechanical Engineering. Vol. 215. 2001, pp. 317-330.

[Le Vie] S. D. Le Vie. Understanding Data Flow Diagrams.
Available: http://ratandon.mysite.syr.edu/cis453/notes/DFD_over_Flow charts.pdf [April 22, 2015].

[Lyadova, 2014] L.N. Lyadova, A.O. Sukhov, E.B. Zamyatina. An Integration of Modeling Systems Based on DSM-Platform. In: Advances in Information Science and Applications. Volumes I & II. Proceedings of the 18th International Conference on Computers (part of CSCC '14). Ed.: E. B. Zamyatina. Vol. 1-2. Santorini Island : CSCC, 2014. P. 421-425.

[Mellor, 2002] S.J. Mellor, M.J. Balcer. Executable UML: A Foundation for Model-Driven Architecture. In: Addison-Wesley Professional, 2002.

[Poryazov, 2009] S. Poryazov. The overlaying free terminal states concept. In: Proceedings of a Conjoint Seminar “Modeling and Control of Information Processes”. 2009, pp. 110-116.

[Poryazov, 2008] S. Poryazov. Towards Useful Overall Network Teletraffic Definitions. In: International Journal “Information Technologies and Knowledge”. 2008, vol. 2, pp. 193-199.

[Poryazov, 2005] S. Poryazov. What is Offered Traffic in a Real Telecommunication Network? In: Proceedings of ITC19/Performance Challenges for Efficient Next Generation Networks. 2005, pp. 707-718.

[Sterle] M. Sterle. Intelligent Assistant for Simulation Model Generation from IDEF3 Process descriptions. Available: http://grantome.com/grant/NSF/IIP-9060443 [June 10, 2015].

[Sukhov, 2013] A.O. Sukhov, L.N. Lyadova. Horizontal Transformations of Visual Models in MetaLanguage System. In: Proceedings of the 7th Spring/Summer Young Researchers’ Colloquium on Software Engineering, SYRCoSE 2013 / Ed.: A. Kamkin.; Ed. by A. Petrenko, A. Terekhov. Kazan, 2013, pp. 31-40.

[Sukhov, 2014, *a*] A.O. Sukhov, L.N. Lyadova. An Approach to Development of Visual Modeling Toolkits. In: Advances in Information Science and Applications. Volumes I & II. Proceedings of the 18th International Conference on Computers (part of CSCC '14). Ed.: E. B. Zamyatina. Vol. 1-2. Santorini Island : CSCC, 2014. P. 61-66.

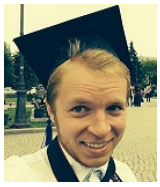
[Sukhov, 2014, *b*] A.O. Sukhov, L.N. Lyadova. Visual Models Transformation in MetaLanguage System. In: Advances in Information Science and Applications. Volumes I & II. Proceedings of the 18th International Conference on Computers (part of CSCC '14) / Ed.: E.B. Zamyatina. Vol. 1-2. Santorini Island, CSCC, 2014. pp. 460-467.

[Zhow, 2015] W. Zhow, F. Yang, Y. Zhu. A transformation method of OPM Model to CPN Model for System Concept Development. In: Proceedings of the First International Conference on Information Science and Electronic Technology (ISET), 2015, pp. 98-102.

[Dorrer, 2011] M.G. Dorrer. Algorithm for Transforming Models of Business Processes into Monochrome Petri Nets. In: Automatic Control and Computer Sciences, Vol. 45, No. 7, 2011, pp. 1-9.

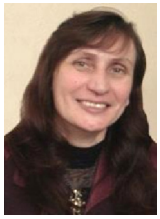
[Lantsev, 2013] Y.A. Lantsev, M.G. Dorrer. Creating agent-based model from the business process discrete-event model. In: St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunication and Control Systems. Vol. 3 (174), 2013, pp. 44-52.

Authors' Information



Roman Nesterov *– National Research University Higher School of Economics, Department of Information Technologies in Business, student; 38, Studenceskaia St., Perm, 614070, Russia; e-mail: RAnesterovHSE@gmail.com.*

Major Fields of Scientific Research: Software Engineering, Business Informatics, Business process analysis, Mathematical modeling, Domain specific modelling



Lyudmila Lyadova *– National Research University Higher School of Economics, Department of Information Technologies in Business, associate professor; 38, Studenceskaia St., Perm, 614070, Russia; e-mail: LLyadova@hse.ru, LNLyadova@gmail.com.*

Major Fields of Scientific Research: Software Engineering, Modelling languages, Visual modelling, Domain specific modelling, Domain specific languages, Language workbenches

POLLEN GRAINS RECOGNITION USING STRUCTURAL APPROACH AND NEURAL NETWORKS

Natalia Khanzhina, Elena Zamyatina

Abstract: *This paper describes the problem of automated pollen grains image recognition using images from microscope. This problem is relevant because it allows to automate a complex process of pollen grains classification and to determine the beginning of pollen dispersion which cause an the allergic responses. The main recognition methods are Hamming network [Korotkiy, 1992] and structural approach [Fu, 1977]. The paper includes Hamming network advantages over Hopfield network [Ossowski, 2000]. The steps of preprocessing (noise filtering, image binarization, segmentation) use OpenCV [Bradsky et al, 2008] functions and the feature point method [Bay et al, 2008]. The paper describes both preprocessing algorithms and main recognition methods. The experiments results showed a relative efficiency of these methods. The conclusions about methods productivity based on errors of type I and II. The paper includes alternative recognition methods which are planning to use in the follow up research.*

Keywords: *image recognition, OpenCV, Hamming network, feature points method, pollen-grains, structural pattern recognition.*

ACM Classification Keywords: *I.5.1 Pattern Recognition Model - Neural nets, Structural, I.5.4 Pattern Recognition Applications - Computer vision*

Introduction

The automated pollen grains recognition problem is a part of rapid-growing and popular intellectual fields - pattern recognition and computer vision. The problem of pollen grains recognition is relevant, as far as the right classification of pollen grains can allow to draw the appropriate conclusions and to solve problems facing biologists (the honey quality control, determination of the pollen dispersion beginning), geologists (determination of the fossil minerals bedding) and experts of other areas. The pollen grains automated recognition problem involves two steps - the image preprocessing and recognition actually.

Today there are some open source software libraries which are used for a particular step of recognition or for some recognition method. One of these libraries is OpenCV. This is the open source computer vision and machine learning software library. The library has more than 2,500 optimized algorithms,

both classic and modern. They can be used for the face recognition and detection, objects identification, movement tracking, and more and, in particular, for the objects classification on an image.

The main steps of pollen grains analysis are collecting (by special-purpose pots), chemical treatment, obtaining images by a microscope, and finally getting statistics [Sladkov, 1967]. At the last stage an expert determines species of plants. Determination of a plant genus is a quite simple task, while the determination of a specie within this genus is often difficult to do. To determine specie of pollen palynologists use a plant atlas, which takes a lot of time. At this stage, the automated recognition of pollen grains would be suitable.

Thus, the recognizing program must be able to determine the amount of pollen grains by input image (a picture obtained by the microscope), localize them on the original image and determine their genuses and species.

The preprocessing step is performed by OpenCV and includes:

1. Image noise reduction, including the feature points method.
2. Image binarization.
3. Pollen grains localization and segmentation.

It was decided to make an attempt of using the structural approach and the neural networks in order to continue the investigations in pollen recognition. The authors [Cherhyh et al, 2013] have tried to use some classical methods of recognition to solve the automated pollen grains recognition problem, but their approach have not produced the desired results, the quantity of correctly recognized images was 54%.

Initial data

A microscope is connected to a computer, so that an expert can get digital photographic images of pollen grains immediately and see them on the computer screen in high quality. As a result of the processing photographic images of pollen grains can be obtained from different sides.

Figure 1 shows images of the same pollen grain from inside and outside. The left image shows inside of the grain. The right image is more blurred, this is the exine, or the outside. The form is less clear, but surface is more clear.

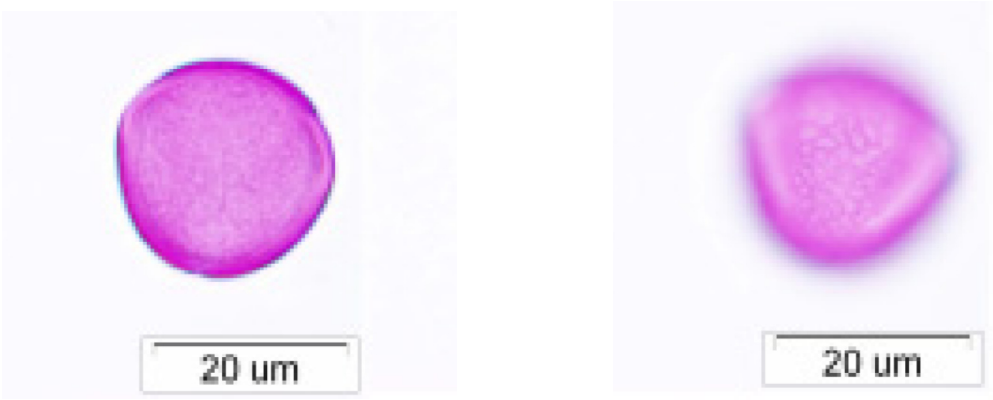


Figure 1. Images of clover pollen grain

Fig. 2 shows the image of the same pollen grain from different angles:

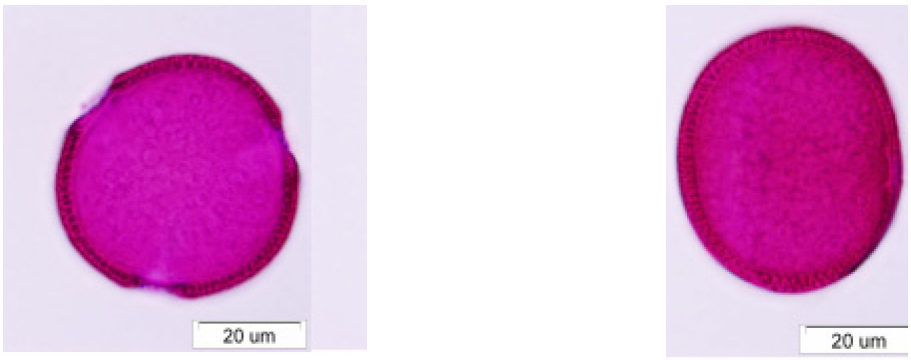


Figure 2. Image of buckwheat pollen grain from different angles

The images may have stains, see Fig. 3:

Figure 3. Image with stains

All these things are causes of complexity of the automated recognition. Therefore it is necessary to perform the preprocessing step in a quality manner. Let's consider the steps of preprocessing: the noise reduction, binarization etc.

Preprocessing

1.1 Noise Reduction

The first operation of noise reduction is smoothing. Smoothing, or blurring, is the simple and frequently used image processing operation. There are many cases where smoothing is needed, but usually it is used to reduce the noise.

The method used for noise reduction is Gaussian blur. Gaussian blur, also known as Gaussian smoothing, is done by convolving each point in the input array with a Gaussian kernel and then summing to produce the output array. The Gaussian filter changes every point by setting its value to the average of all points in some radius (corresponding to the kernel of smoothing).

The following steps of noise removing are using such morphological operations as the dilation and erosion functions.

The erode operation is often used to eliminate "speckle" noise in an image. The idea here is that the speckles are eroded to nothing while larger regions that contain visually significant content are not affected. The dilate operation is often used when attempting to find connected components. The utility of dilation arises because in many cases a large region might otherwise be broken apart into multiple components as a result of noise, shadows, or some other similar effect. A small dilation will cause such components to "melt" together into one [Bradsky et al, 2008].

A good example of using of these operations is shown in Figure 4:

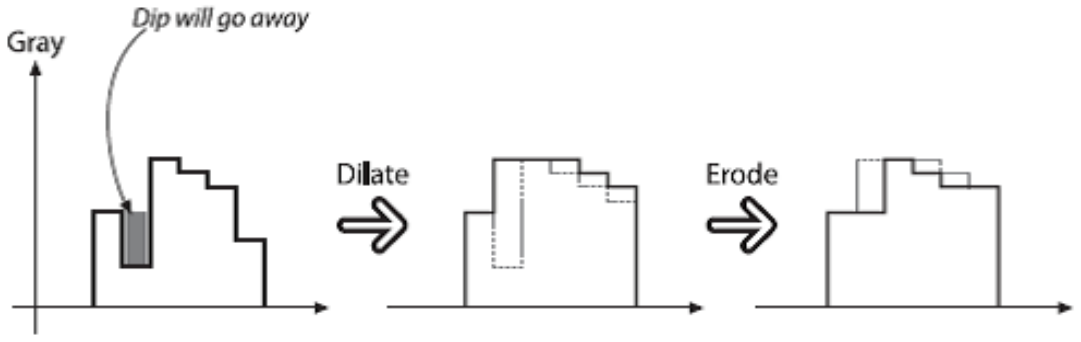


Figure 4. Using of dilation and erosion

Repeated using of these functions on the same image can give a significant noise reduction as a result.

1.2 Binarization

The next step of the preprocessing is the image thresholding.

To perform this step, the following algorithm is applied:

1. The image is converted from the RGB color model (Red, Green, Blue) to the HSB color model (Hue, Saturation, Brightness).

2. All pixels of the necessary hue (from claret to violet, the dye used in a treatment of pollen always has a color of that hue) are converted to the black. This kind of binarization is double-threshold according to the following formula (1):

$$f'(m,n) = \begin{cases} 0, f(m,n) \geq t_1; \\ 1, t_1 < f(m,n) \leq t_2; \\ 0, f(m,n) > t_2, \end{cases} \quad (1)$$

where t_1, t_2 - threshold values, f - the input image, f' - the output image, m,n - pixel coordinates, $t_1<t_2$.

3. Saturation of all the pixels whose saturation value is more than 30 (where maximum is 255) is maximized, saturation of the rest is minimized. Thus, this is the low-threshold binarization.

This operation is a quiet simple and uses the only one threshold value according to the formula (2):

$$f'(m,n) = \begin{cases} 255, f(m,n) \geq t \ ; \\ 0, f(m,n) < t \ , \end{cases} \quad (2)$$

where t - threshold value,f - the input image, f '- the output image, m, n - pixel coordinates.

4. Calculating the conjunction of the hue and saturation layers.

Let's consider the next step - segmentation.

1.3 Segmentation

The goal of segmentation in this case is to separate pollen grains contained on the image.

One of the stages of segmentation is contour finding. OpenCV has several methods for this, one of them is the Canny edge detector.

The most significant dimension to the Canny algorithm is that it tries to assemble the individual edge candidate pixels into contours. These contours are formed by applying an hysteresis threshold to the pixels. This means that there are two thresholds, an upper and a lower. If a pixel has a gradient larger than the upper threshold, then it is accepted as an edge pixel; if a pixel is below the lower threshold, it is rejected. If the pixel's gradient is between the thresholds, then it will be accepted only if it is connected to a pixel that is above the high threshold [Bradsky et al, 2008].

The found edges can help to find the contours of pollen grains. Then recursive algorithm builds from the contours tree a list of those that can be pollen grain contour:

1. The first stage is to remove too small areas of contours.
2. The areas containing contour is compared with pollen grains patterns using Hu-moments.
3. The rest of contours are filled with black color. These are the separate pollen grains already.

There are the results of the preprocessing (Fig. 5-7):

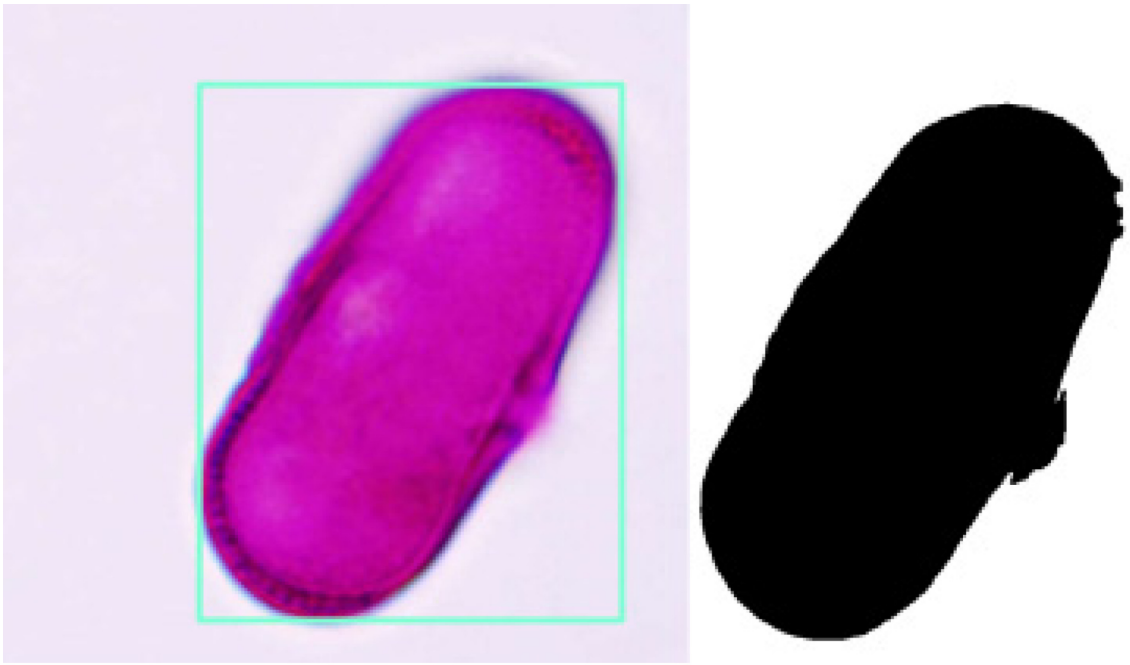


Figure 5.The result of angelica image preprocessing

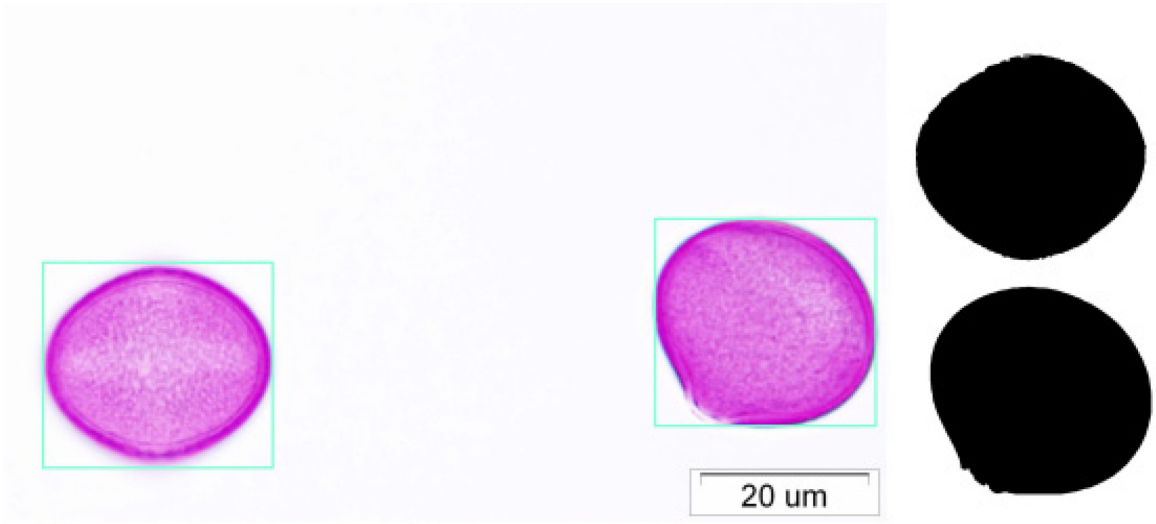


Figure 6. The result of clover image preprocessing

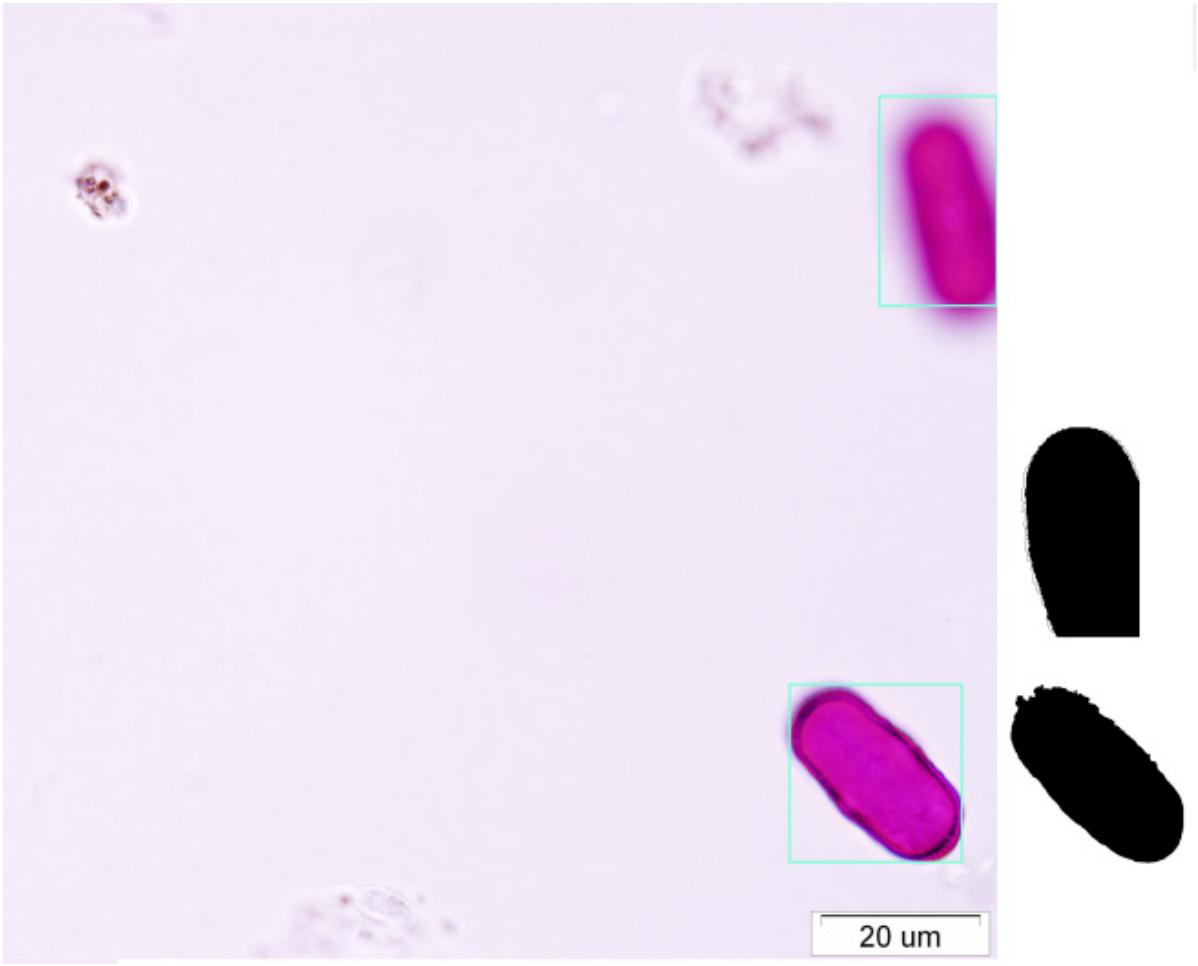


Figure 7. The result of preprocessing of angelica image with stains

1.4 Scaling

For those recognition methods which are not invariant to image scale it needs to apply another stage of the image preprocessing called scaling.

All the images are scaled to the size of 200x200. That allows, for example, to compare the input images with patterns using Hamming metric (an amount of distinct bits in a two-dimensional vector representing an image).

Image recognition methods

1.5 Structural approach

This method is also called structural, because an object is described as a grammar, not as a features vector. The recognition presents parsing.

The adaption of this approach to two-dimensional objects like images is difficult. Firstly, there is no obvious way of choice of finite elements. Secondly, there is no obvious way of choice of reduction rules [Fu, 1977].

The alternative for the structural approach is the Freeman chain code. Finite elements are the numbers from 0 to 7. The encoder moves along the boundary of the object and, at each step, transmits a symbol representing the direction of this movement (Fig. 8).

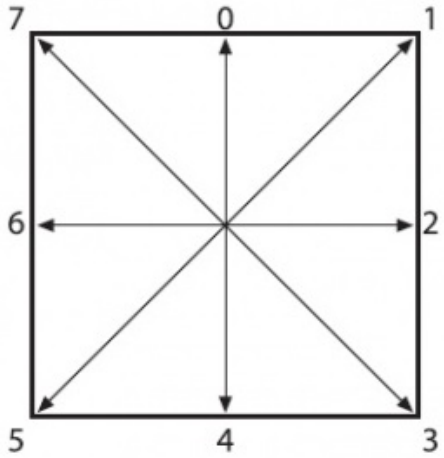


Figure 8. Directions of movement

Thus, the Freeman code is a sequence of numbers describing the boundary of the image. An example of such code is shown in Figure 9, the entry arrow is red.

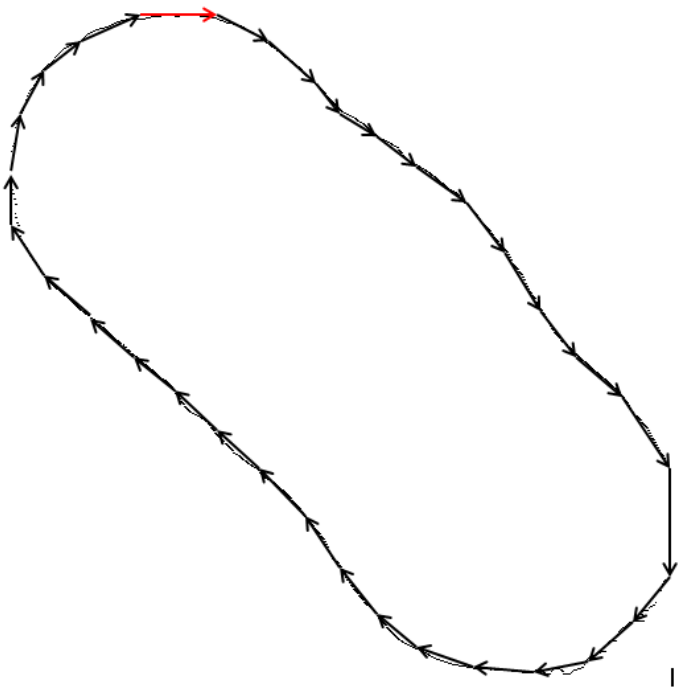


Figure 9. Chain of the pollen grain image

Fig. 10 shows how boundary looks in a program:

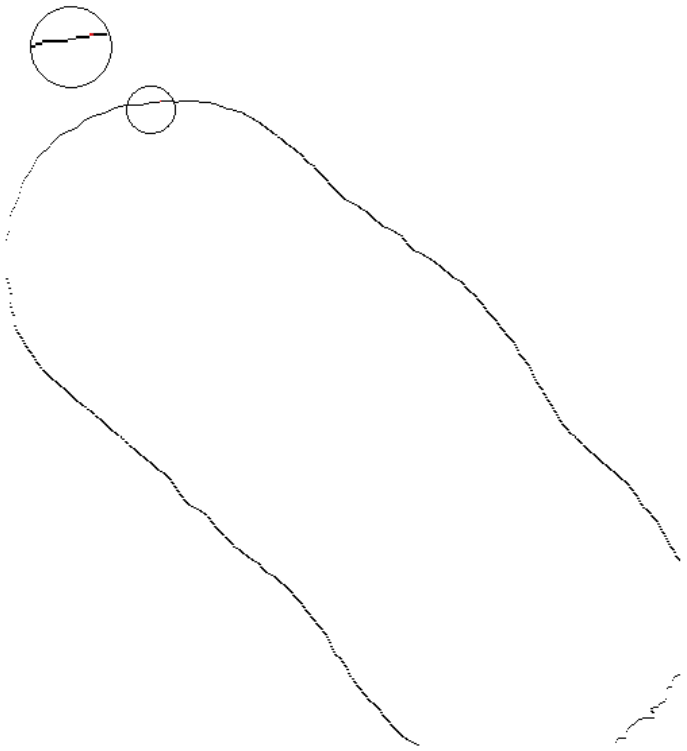


Figure10. The boundary of pollen grain image in program

Next, the chain code of the image is compared with the chains of patterns, the closest one is the answer (in sense of Euclidean distance). Consider the next method of recognition.

1.6 Hamming neural network

The usage of a neural network is one of non-classical methods of pattern recognition. The training set here is the set of patterns including each plant's specie, from different angles of pollen grains.

The chosen architecture is Hamming network. This network requires a few memory and small amount of computation, against, for example, Hopfield network. Figure 11 shows Hamming neural network layout.

The advantage of that network is a small amount of weighted connections between neurons. Hopfield network with input size of 100 can memorize 10 patterns, while its layout will have 10,000 synapses. The Hamming network layout with the same capacity will have only 1000 synapses [Ossowski, 2000].

The Hamming network consists of two layers. The first and second layers each have m neurons, where m is the training set capacity. The first layer neurons have n synapses connected to the input (which is called zero layer). The second layer neurons are linked by negative synaptic feedback. For each neuron the only one synapse with positive feedback is connected to his own axon.

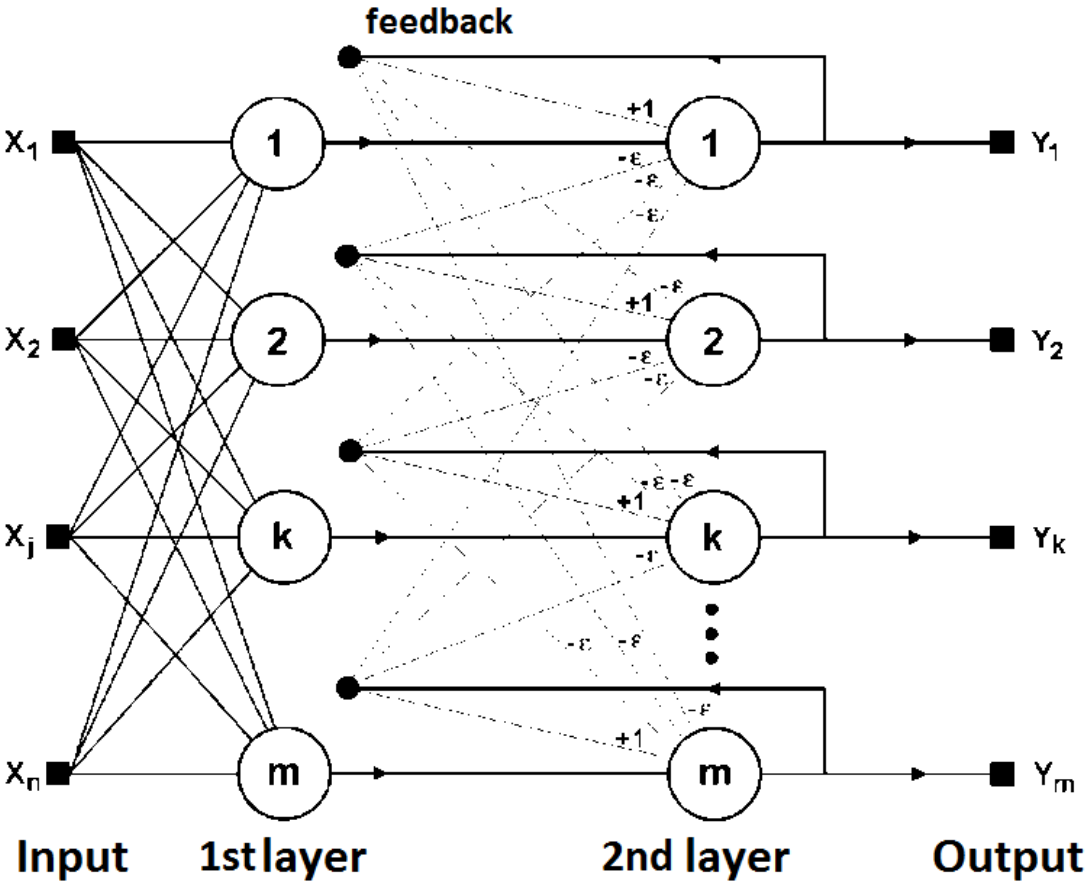


Figure11. Hamming neural network layout

The idea of the network is to compute the minimum of Hamming distances between an input image and all patterns. The network should activate the only one output corresponding to the pattern with the minimum distance.

At the stage of initialization weights of the first layer and the activation function threshold have the following values:

$$w_{ik} = \frac{x_i^k}{2},\ i=0...n-1,\ k=0...m-1 \tag{3}$$

$$T_k = n / 2,\ k = 0...m-1 \tag{4}$$

where x_i^k - i-th element of the k-th pattern.

The weights of inhibiting synapses in the second layer have a value between 0 and 1/m. A synapse neuron which is connected with his own axon has a weight of 1.

The Hamming network has the following algorithm [Korotkiy, 1992]:

1. The input is an unknown vector $\boldsymbol{X} = \{x_i: i=0 \ldots n-1\}$, the first layer neurons state is calculated from the input (the superscript means the layer number):

$$y_j^{(1)} = s_j^{(1)} = \sum_{i=0}^{n-1} w_{ij} x_i + T_j,\ j=0...m-1 \tag{5}$$

The resulting values are the inputs to the axons of the second layer:

$$y_j^{(2)} = y_j^{(1)},\ j = 0...m-1 \tag{6}$$

2. The second layer neurons new state calculation:

$$s_j^{(2)}(p+1) = y_j(p) - e\sum_{k=0}^{m-1} y_k^{(2)}(p), k \neq j,\ j = 0...m-1 \tag{7}$$

and their axons values calculation:

$$y_j^{(2)}(p+1) = f\left[s_j^{(2)}(p+1)\right],\ j = 0...m-1 \tag{8}$$

The activation function f is a threshold, the F value must be large enough to avoid the oversaturation.

3. Check if the second layer outputs changed since the last iteration. If yes then go to the step 2. Else this is the end.

The only problem with Hamming neural network is when images with noise are at the same Hamming distance from two or more patterns. In this case, the choice between these patterns is random [Ossowski, 2000].

1.7 The feature points method

The following image recognition method is the feature points method.

The image is represented as a set of special key points. A scene feature point or a point feature is a point of the image, whose neighborhood is stand out from other points neighborhood [Bay et al, 2008].

A feature point can be determined as a point of a sudden gradient drop in the image by two directions (corner points).

The feature points are compared with each other by their descriptors. There is the class in OpenCV for describing the feature points named SURF (Speeded up Robust Features). This is one of descriptors, which searches feature points and describes them in invariant to the scaling and rotation way. Besides, the feature points are invariant in the sense that the object on the image from different points of view has the same set of feature points [Bovyrin et al, 2013].

The feature points method does not give good results: only 35% of the right classification. On Fig. 12 feature points comparison gave the right answer, in Fig. 13 - the wrong answer.

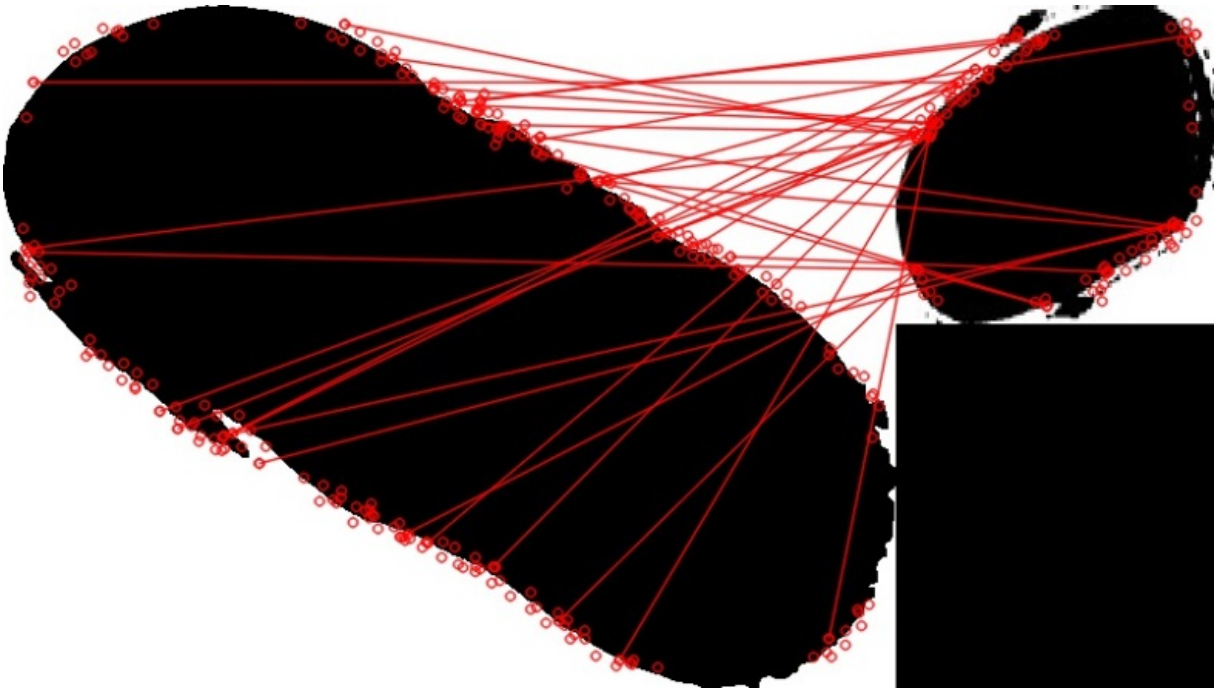


Figure 121. Example of correct recognition by the feature points method

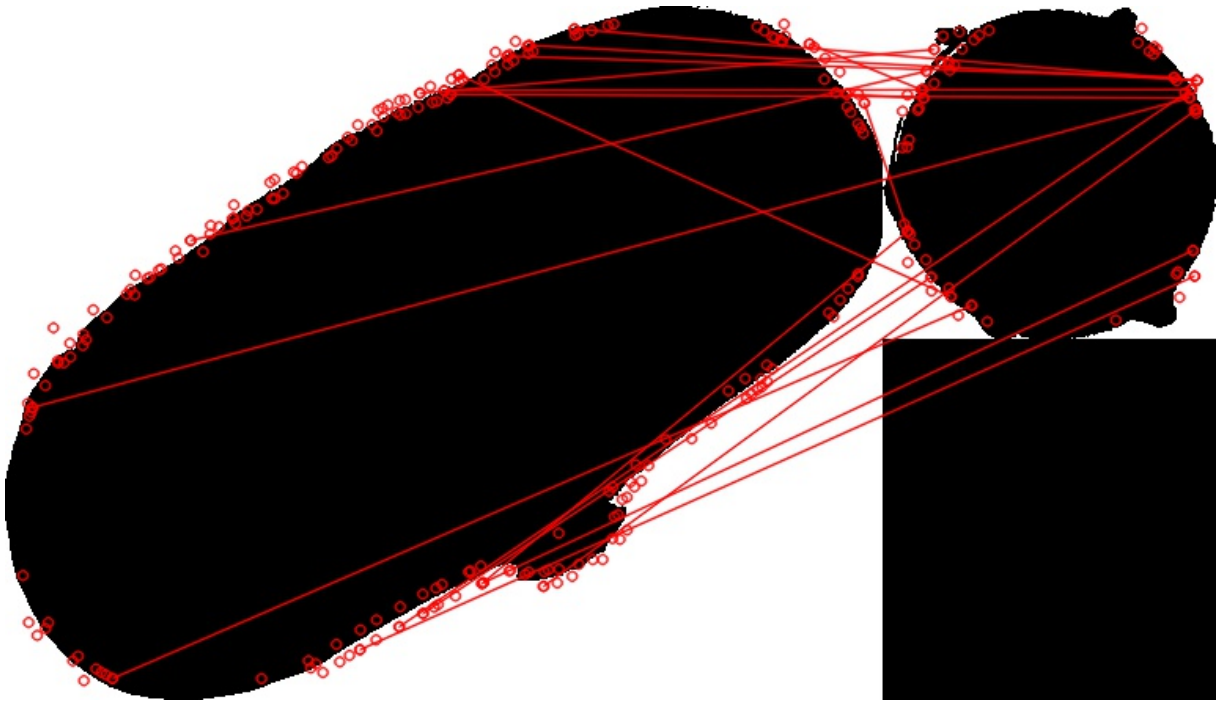


Figure13. Example of incorrect recognition by feature points method

The feature points method was less effective in the pollen grains images classifying, but good enough to throw out stains, which are not pollen grains, but have the same color. Perhaps the cause of low efficiency is the rounded shape of pollen grains of different species. Thus, this method completes the preprocessing.

Results

The estimation of efficiency of all methods is based on measure of errors of the first and second kinds. The error of the first kind is the incorrect miss (false negative), when an event of interest is not detected. The error of the second kind is incorrect detection (false positive), when an event of interest is absent, but is detected [Vezhnevec, 2007].

Let us consider the neural network estimation results (Table. 1):

Table 1. Results of testing the neural network

	Angelica	Clover	Buckwheat	Pigweed	Carnation	Averages
Pollen grains amount	122	135	53	74	73	452
Normalized rate of errors of the first kind	20%	36%	26%	23%	32%	28%
Normalized rate of errors of the second kind	28%	5%	0%	0%	0%	7%
Percent of correct misses	72%	95%	100%	100%	100%	93%
Percent of correct detections	80%	64%	74%	77%	68%	72%

The structural method gave the highest result of angelica recognizing: 44% of errors of the first kind, 0% of errors of the second kind, 100% of correct misses and 56% of correct detections. The average of correct detections across five species is 42%.

The combination of all three methods gave a good result in the sense of errors of the second type, the average is 7% for the neural network, and 0% for the structural method.

The average of correct detections is 72% for the neural network and 42% for the structural method.

The cause of the most of the errors of the first kind is the pollen grains taken from the exine.

Conclusion

The feature points method is well suited for exclusion from the list for recognition those objects which are not pollen grains, like stains. Hamming neural network and structural approach do not distinguish stains from grains, this is their main disadvantage. The combination of preprocessing basic methods with the feature points method maximizes stains elimination for further recognition by the neural network.

This approach has the following results: for the neural network an average error rate of the first kind is 27%, of the second kind - 7%. Accordingly, neural network recognized correctly about 72% of the pollen grains. Structural method detected correctly about 42% of the pollen grains.

The cause of the most of the first kind errors is the pollen grains taken from the exine, that is, their boundaries are blurry. The next problem is the pollen grains stuck together, but this happens rare, about two pairs per 250 grains.

The next step in research is to use the OpenCV's boosting. It is also planned to apply a texture recognition, a support vector machine. Research of neural networks is not finished: the next network would be convolutional neural network.

One of the major drawbacks of the used methods is that the speed of processing and recognition is quite low. It takes about 40 seconds per an image with one pollen grain. Obviously, the program needs an optimization. Therefore, one of the following stages of development is the usage of concurrency.

Bibliography

[Sladkov, 1967] *Sladkov A.* Introduction to the spore-pollen analysis. — M.: 1967 (in Russian)

[Cherhyh et al, 2013] *Cherhyh A., Zamyatina E.* Problems of an application of the classical methods of pattern recognition for the photographic images of pollen grains. Proceeding of the Conference "Image analysis, networks and texts" (AIST'2013), Ekaterinburg, Russia, 4-6 of April 2013 r., M: The National Open University "INTUIT", 2013, ISBN 978-5-9556-0148-9, pp. 160-168. (in Russian)

[Bradsky et al, 2008] *Bradsky G., Kaehler A.* Learning OpenCV. Computer Vision with the OpenCV Library, 2008.

[Fu, 1977] *K.S. Fu* Syntactic Pattern Recognition and Applications/ Springer - Berlin, 1977.

[Ossowski, 2000] *S. Ossowski,* Neural networks for information processing, Of. Ed. Pol. Warsaw, Warsaw, Poland 2000in Polish. pp. 176-186

[Korotkiy, 1992] *Korotkiy S.* Hopfield and Hamming Neural networks/ Internet: URL: http://www.shestopaloff.ca/kyriako/Russian/Artificial_Intelligence/Some_publications/Korotky_Neuron_network_Lectures.pdf, pp. 56-59.

[Bay et al, 2008] *Bay H, Ess A., Tuytelaars T., Van Gool L.* Speeded-Up Robust Features (SURF)/ Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, , 2008 - Zurich,Switzerland; Leuven, Belgium, pp. 346--359

[Bovyrin et al, 2013] *Bovyrin A., Druzhkov P., Eruhimov V.* Key points detectors and descriptors. Algorithms for image classification. The problem of detecting objects in the images and the methods of its solutions Internet: URL:http://www.intuit.ru/studies/courses/10621/1105/lecture/17983?page=2

[Vezhnevec, 2007] *Vezhnevec V.* The estimation of quality of classifiers // Online Journal "Computer graphics and multimedia"/ Internet: URL: http://cgm.computergraphics.ru/content/view/106

Authors' Information

Natalia Khanzhina *– Dept. Informational technologies and programming, Saint Petersburg National Research university of information technologies, mechanics and optics, Russia, 197101, St. Petersburg, 49 Kronverksky Pr.; e-mail: nehanzhina@gmail.com*

Major Fields of Scientific Research: Image recognition, Neural networks

Elena Zamyatina *– Researcher; Perm State National Research University, Russia, 614990, Perm, Bukireva st., 15; e-mail: e_zamyatina@mail.ru*

Major Fields of Scientific Research: Pattern recognition, Simulation

ОСОБЕННОСТИ АНАЛИЗА СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ В СФЕРЕ ИНФОКОММУНИКАЦИЙ

Галина Гайворонская, Петр Яцук, Юлия Казак

Аннотация: *Огромные объемы статистической информации, которая стала доступной при построении информационного общества, привели к увеличению случаев некорректного использования методов анализа этой информации, что зачастую влечет за собой ошибочные выводы, которые могут привести к негативным последствиям. Во избежание такой ситуации в работе даны некоторые рекомендации по условиям применения методов анализа данных и сформирована методика поэтапного анализа статистки в сфере электронных коммуникаций.*

Ключевые слова*: статистическая информация, анализ данных, информационно-коммуникационные технологии, инфокоммуникационные услуги.*

ACM Classification Keywords*: G.3 Mathematics Of Computing – Probability And Statistics – Statistical Computing, Time Series Analysis, Correlation and Regression Analysis; C.2 Computer-Communication Networks.*

Введение

В настоящее время в процессе перехода от индустриального общества к информационному невозможно переоценить роль всеобъемлющей и достоверной информации о происходящих в обществе процессах и явлениях. Естественно такую информацию целесообразно получать путем анализа имеющейся статистической информации, однако в последние годы результаты статистических отчетов все чаще вызывают недоверие, связанное с неадекватным отображением ими действительности. Это связанно не только с заведомой подтасовкой и искажением имеющейся статистической информации, но зачастую и с неумением правильно применять методы анализа данных.

Если в предыдущем столетии анализом статистики, связанной с передачей информации по информационным сетям, занимались в основном специалисты, осуществляющие весь комплекс мероприятий, связанных с получением, (чаще всего путем специально организованных измерений) и обработкой статистической информации, то теперь, когда информация и

программные средства ее обработки стали более доступными, ее анализом пытаются заниматься и дилетанты, плохо представляющие условия применения тех или иных методов анализа данных (АД). Такая ситуация приводит к неверным результатам на основании которых делаются ошибочные выводы, иногда приводящие к далеко идущим последствиям не лучшим образом, сказывающимся на развитии нашей страны.

Еще одной особенностью сложившейся ситуации стало то, что если раньше специалисты в основном сталкивались с нехваткой достоверной статистической информации, то сейчас для получения несмещенных оценок исследуемой случайной величины необходимо обрабатывать сверхбольшие объемы статистики. Эти объемы измеряются десятками, а то и сотнями гигабайт, что тоже вызывает определенные трудности, так как ни один программный продукт, предназначенный для обработки статистики не способен корректно обработать такие большие массивы данных.

Следующая сложность связана с наличием большого количества специализированных программных продуктов, предназначенных для анализа данных (ППАД). Их наличие иногда, вместо того, чтобы помочь специалистам, сокращая трудозатраты на выполнение трудоемких рутинных операций, приводит к некорректному использованию, при котором исследователь зачастую даже не понимает, что именно рассчитывается ППАД, по каким формулам и при каких условиях допустимо использование тех или иных методов АД, что априори приводит к заведомо неверным результатам, а соответственно и к ошибочным выводам, сделанным на основе этих результатов.

Постановка задачи

Все это, также и ряд других немаловажных аспектов привели к необходимости разработки методики корректного применения методов анализа данных в инфокоммуникациях. Одному из подходов решения такой задачи посвящена эта работа. Существует большое количество учебной литературы и научных исследований, в которых дается определение тех или иных методов анализа статистической информации. К наиболее известным из них можно отнести работы Андерсона Т. [Андерсон, 1963; Андерсон, 1976], Айвазяна С.А. [Айвазян, 1998; Айвазян, 1989], Боровкова А.А. [Боровков, 2010], Горяинова В.Б. [Горяинов, 2001], Иберлы К. [Иберла, 1980], Кобзаря А.И. [Кобзар, 2006], Орлова А.И. [Орлов, 2004], Мендаля И.Д. [Мандель, 1988], Ким Дж. О. [Ким, 1989], Лоули Д. [Лоули, 1967], Окунь Я. [Окунь, 1974], Хармана Д. [Харман, 1972], Ферстера Э. [Ферстер, 1983], Ренца Б. [Ферстер, 1983], Кендалла М. [Кендалл, 1976], Стьюарта А. [Кендалл, 1976], Лемана Е. [Леман, 1964], Ллойда Э. [Ллойд,1989; Ллойд, 1990], Ледермана В.

[Ллойд,1989; Ллойд, 1990], и т.д. Однако большинство из них позволяют хорошо изучить теорию, но не дают практических рекомендаций по получению навыков применению этих методов и использованию для этих целей, имеющихся ППАД.

Наиболее полно задача использования методов АД и существующих ППАД отражена в работах проф. Н.Д. Вайсфельд [Вайсфельд, 2006], которая на протяжении многих лет консультировала сотрудников кафедры информационно-коммуникационных технологий Одесской национальной академии пищевых технологий (ИКТ ОНАХТ) в процессе диссертационных исследований, связанных с применением методов АД, при решении разнообразных задач, связанных с анализом и синтезом информационных сетей Украины. Так в частности результаты этих исследований опубликованные в работах [Гайворонская, 2009a; Гайворонская, 2007a; Гайворонская, 2007b; Павлов, 2007; Котова, 2009a; Котова, 2010a; Котова, 2010b; Котова, 2008; Котова, 2009b; Сахарова, 2008; Бондаренко, 2010a; Бондаренко, 2010b; Бондаренко, 2011], дали возможность обобщить некоторые особенности применения методов АД при исследованиях в области инфокоммуникаций.

Так, например, в работах [Гайворонская, 2009b; Ганницкий, 2009; Гайворонская, 2009c; Ганницький, 2011; Гайворонская, 2007c; Ганницкий, 2008a; Ганницкий, 2008b; Gannitskiy, 2010; Ганницкий, 2010a; Ганницький, 2010b; Gannytskyi, 2012], опубликованных по результатам диссертационных исследований И.В. Ганницким на основе анализа методов обработки статистической информации показано, что применение существующих статистических программных продуктов не дает возможности обрабатывать сверхбольшие массивы данных объемом более 1 млн. записей. Поэтому им разработан метод аналитической обработки статистических данных, включающий кеширование, семплинг, иерархическое структурирование исходной информации, что позволило обработать сверхбольшие объемы данных и повысить точность и скорость обработки для исследования массивов статистических данных сверхбольшого объема (более 250 ГБ).

Обработка результатов измерений, выполненная с использованием авторского программного продукта, позволила проанализировать тенденции изменения параметров потоков вызовов и поступающей нагрузки. В результате показано, что существенным является явно выраженное снижение темпов роста нагрузки операторов мобильной связи, количество вызовов мобильной связи и их длительность имеют ярко выраженную тенденцию падения, а пики нагрузки в

праздничные дни становятся все более ярко выраженным при общей тенденции сглаживания нагрузки.

Разработанный автором метод повышения эффективности обработки сверхбольших объемов исходных данных с использованием системы управления базой данных (СУБД) *Oracle*, отличающийся использованием специфических внутренних функций СУБД, переносом методов статистической обработки на уровень работы СУБД позволяет:

- Сократить время, необходимое для обработки статистических данных;
- Повысить скорость доступа к обработанным данным в 6–7 раз;
- Автоматизировать создание отчетов в различных временных срезах и режимах;

Увеличить точность полученных результатов на 12-15% за счет подбора интервала усреднения и обеспечить взаимодействие с существующими программными продуктами, предназначенными для обработки статистики.

Можно утверждать, что в большинстве исследований, связанных с развитием сферы инфокоммуникаций в той или иной степени присутствуют результаты, полученные с применением методов АД.

Эта работа направлена на разработку пошагового алгоритма, определяющего условия корректного применения методов АД и рекомендаций по использованию ППАД, наиболее удобных для решения отдельных задач АД, относящихся к анализу функционирования и планированию развития информационных сетей и информационного общества в целом.

Основная часть

Прежде чем приступить к использованию методов математической статистики необходимо определить способ представления имеющейся статистической информации, которая может представлять собой выборку или временной ряд. Если имеющаяся статистическая информация упорядочена во времени - она представляет собой временной ряд, если же сформирована по каким-то другим признакам, ее называют выборкой. В литературе, посвященной использованию методов математической статистики [Кобзар, 2006] под выборкой понимают результат n независимых последовательных наблюдений случайной величины X из рассматриваемой генеральной совокупности.

Для того чтобы определить какие именно методы корректно использовать для анализа конкретной выборки, необходимо оценить ее вид, то есть определить является ли выборка стандартной, парной, комбинированной, стратифицированной, кластерной или систематической.

— Если используется парная статистическая выборка, то прежде, всего необходимо оценить является ли она репрезентативной, для этого она должна содержать не менее 30 объектов, в случае, если выборка содержит меньшее количество объектов, использование статистических методов для ее исследования считается некорректным.

— Если стандартная статистическая выборка репрезентативна, проводится анализ характера распределения исследуемой случайной величины.

— Если выборка подчиняется хотя бы одному из распределений известного вида, то ее можно исследовать методами параметрической статистики, при этом определяются основные центральные моменты, а именно математическое ожидание, дисперсия и среднеквадратическое отклонение.

— Если вид функции распределения исследуемой случайной величины не может быть описан известными законами распределения, то расчет перечисленных основных центральных моментов не имеет смысла. В этом случае необходимо использовать методы непараметрической статистики, к которым относится вычисление: медианы, то есть числа, которое делит вариационный ряд на две части, содержащие одинаковое число элементов, если объем выборки $n = 2l + 1$ (непарный), то $Me_x = x^{(l+1)}$, если же объем выборки парный, то $Me_x = \frac{(x^{(l+1)} + x^{(l+1)})}{2}$; моды – величины признака, чаще всего встречающегося в данной совокупности $\mathrm{M_o} = x_0 + n\frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})}$; размаха – разницы между максимальным и минимальным значением признака $R = x_{\max} - x_{\min}$ и квартиля – значения признака в ранжированном ряду распределения, выбранного таким образом, что 25% единиц совокупности будут меньше по размеру $\mathrm{Q_1}$; 25% будут заключены между $\mathrm{Q_1}$ и $\mathrm{Q_2}$; 25 % – между $\mathrm{Q_2}$ и $\mathrm{Q_3}$; другие 25 % превосходят $\mathrm{Q_3}$:

$$\mathrm{Q_1} = x_{Q_1} + i_{Q_1}\frac{\frac{1}{4}\sum f - S_{\mathrm{Q_1}-1}}{f_{Q_1}},\ \mathrm{Q_2} = x_{Q_2} + i_{Q_2}\frac{\frac{2}{4}\sum f - S_{\mathrm{Q_2}-1}}{f_{Q_2}},\ \mathrm{Q_3} = x_{Q_3} + i_{Q_3}\frac{\frac{3}{4}\sum f - S_{\mathrm{Q_3}-1}}{f_{Q_3}}.$$

На Рисунке 1 представлен общий подход к использованию методов математической статистики.

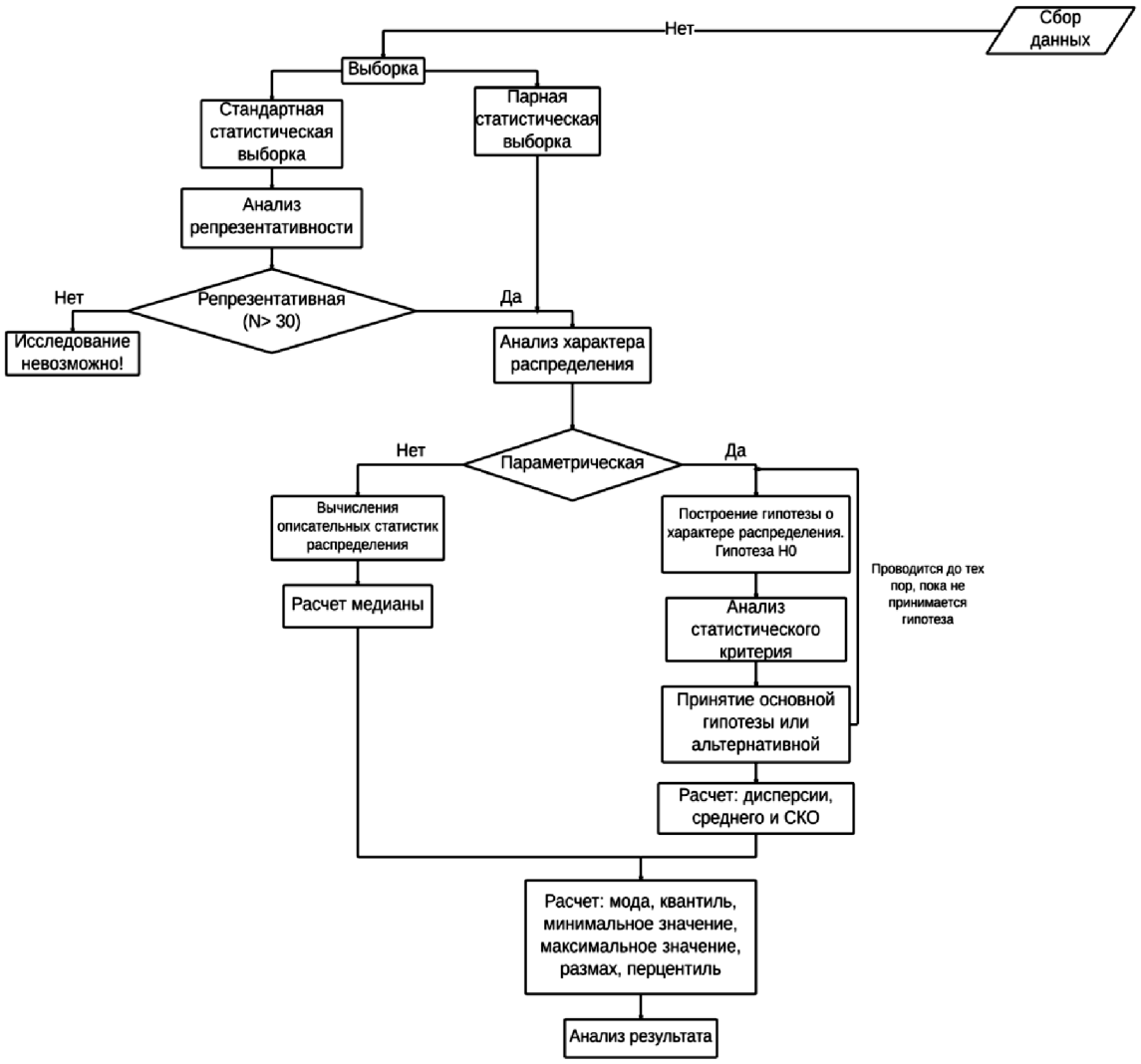


Рисунок 1. Методы анализа статистической информации для выборки

В качестве примера предварительного анализа статистической информации рассмотрим результаты, полученные при анализе состояния телекоммуникационной отрасли в различных странах мира, размещенной на сайтах Международного союза электросвязи (*International Telecommunication Union, ITU)* [ITU] и Европейской экономической комиссии ООН (*United Nations Economic Commission for Europe, UNECE)* [UNECE Statistic Database]. На основе этих данных выполнен анализ информации, характеризующей количество пользователей четырех видов связи: фиксированной проводной широкополосной связи (ФПШС), обозначенной T1, фиксированной телефонной связи (ФТС) (T2), мобильной связи (МС) и Интернет (T4). В

исследовании использованы показатели для 53 стран мира за период с 2000 по 2012 годы. Пример анализируемых исходных данных приведен в Табл. 1 - 4, отражающих использование ФПШС, ФТС, МС и процент пользователей Интернет.

Таблица 1. Количество точек подключения T1

Страна	**2000**	**2001**	**2002**	…	**2009**	**2010**	**2011**	**2012**
Австрия	190500	320600	451000	…	1845600	2029000	2078000	2118000
Молдова		239	418	…	186973	269067	355099	417177
Россия	0	0	11000	…	12900000	15700000	17420161	20703653
США	7069874	12792812	19881549	…	78349000	82759000	86445000	90341000
Узбекистан		0	0	…	88735	118000	147760	212729
Украина			0	…	1906725	2954556	3169396	3643460
Швейцария	56416	140000	455220	…	2739149	2911504	3076384	3210631

Таблица 2. Количество точек подключения T2

Страна	**2000**	**2001**	**2002**	…	**2009**	**2010**	**2011**	**2012**
Австрия	3997000	3997000	3883000	…	3253000	3398000	3388000	3380000
Молдова	583811	639165	719286	…	1138729	1161148	1179953	1205768
Россия	32070000	33278200	35500000	…	45379601	44915829	44151461	42168388
США	192513000	191570800	189250143	…	152873000	149652000	143319000	138595000
Узбекистан	1655044	1662963	1681127	…	1856592	1892164	1927735	1980038
Украина	10417000	10669600	10833300	…	13026293	12941346	12680881	12182142
Швейцария	5235733	5383483	5387568	…	5131810	4907773	4898770	4721981

Таблица 3. Количество точек подключения Т3

Страна	**2000**	**2001**	**2002**	**…**	**2009**	**2010**	**2011**	**2012**
Австрия	8562000	11132000	12670000	…	22200000	22500000	24490000	24338000
Молдова	139000	225000	338225	…	2784832	3165052	3587431	4080143
Россия	3263200	7750499	17608756	…	230050000	237689224	256116581	261886329
США	109478031	128500000	141800000	…	274283000	285118000	298293000	310000000
Узбекистан	53128	128012	186900	…	16417914	20952000	25441789	20274090
Украина	818524	2224600	3692700	…	54942815	53928830	55576481	59343693
Швейцария	4638519	5275791	5736303	…	9322580	9644157	10082636	10460000

Таблица 4. Количество точек подключения Т4

Страна	**2000**	**2001**	**2002**	**…**	**2009**	**2010**	**2011**	**2012**
Австрия	33,73	39,19	36,56	…	73,45	75,17	78,74	80,00
Молдова	1,28	1,49	3,79	…	27,50	32,30	38,00	43,37
Россия	1,98	2,94	4,13	…	29,00	43,00	49,00	63,80
США	43,08	49,08	58,79	…	71,00	71,69	69,73	79,30
Узбекистан	0,48	0,60	1,08	…	17,06	20,00	30,20	36,52
Украина	0,72	1,24	1,87	…	17,90	23,30	28,71	35,27
Швейцария	47,10	55,10	61,40	…	81,30	83,90	85,19	85,20

Из Таблиц 1 - 4 видно, что имеющаяся статистическая информация представляет собой стандартную статистическую выборку, состоящую из 53-х объектов, что подтверждает ее репрезентативность. При анализе характера распределения выборка сводится к однородному

показателю. Случайная величина, подчиняющаяся нормальному распределению на графике, имеет вид кривой „колоколообразной" формы, в пределах которой находятся все ее значения.

На Рис. 2 представлены гистограммы распределения случайной величины, представленной в анализируемой выборке с помощью критерия χ -квадрат.

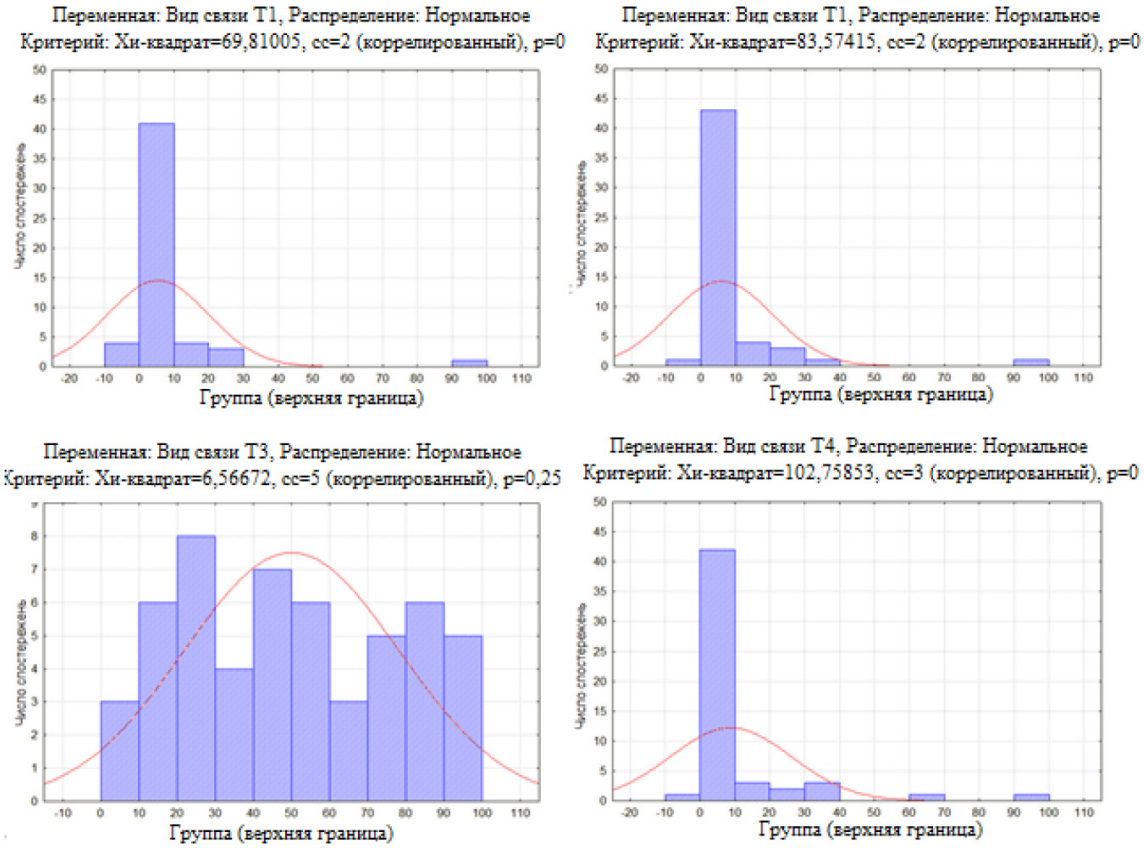


Рисунок 2. Гистограммы анализа нормального распределения для видов связи T1 - T4

Для достоверности результатов проверки нормального распределения выполнено исследование тех же выборок с использованием еще трех критериев:

- Колмогорова-Смирнова основанного на максимуме разности между кумулятивным распределением выборки и предполагаемым кумулятивным распределением;
- Лиллиефорса представляющего модификацию критерия Колмогорова-Смирнова;
- *W*-критерия Шапиро-Уилки являющегося наиболее эффективным, так как он обладает большей мощностью по сравнению с альтернативными критериями проверки нормальности.

Результаты продемонстрированы на Рис. 3.

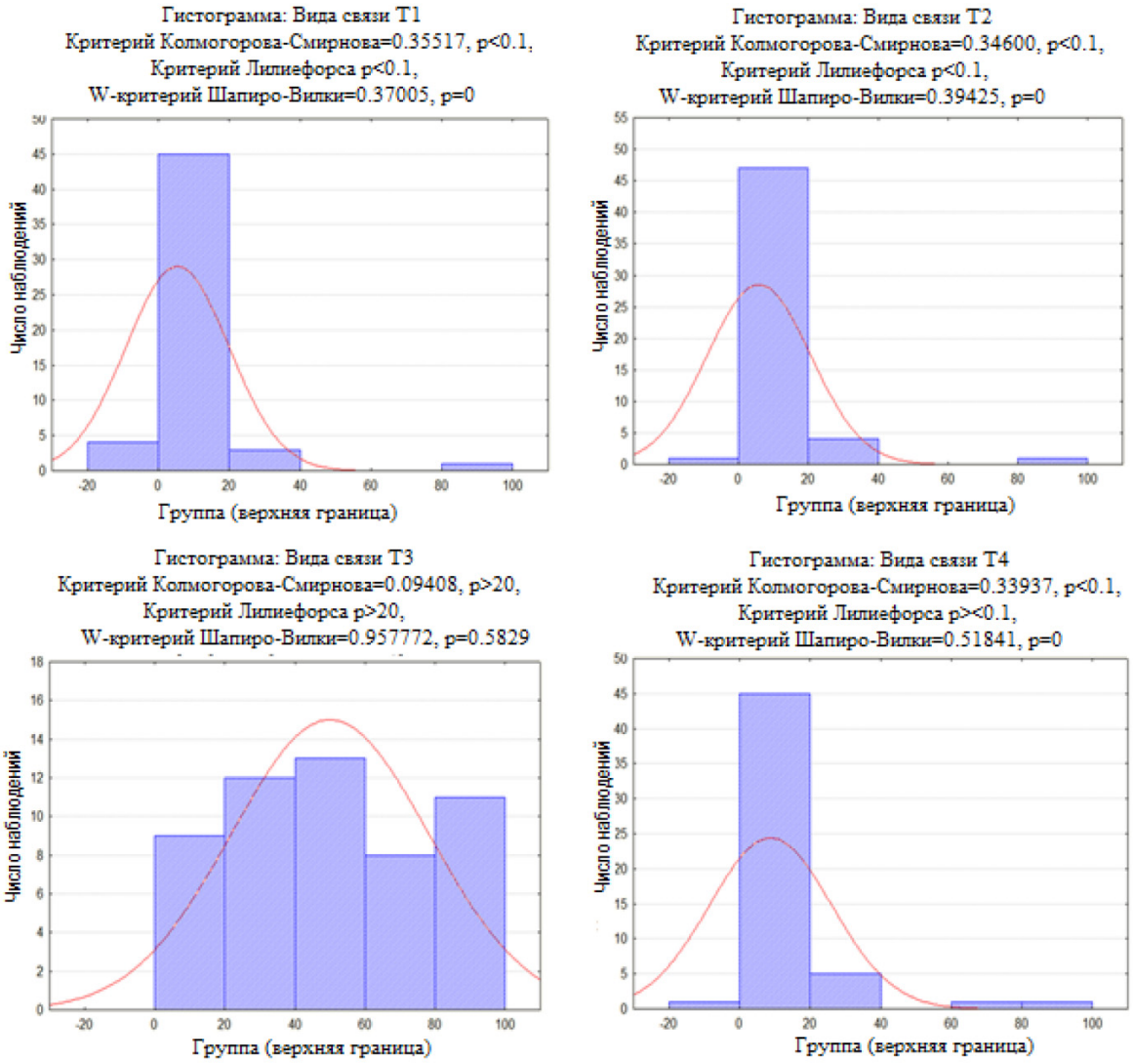


Рисунок 3. Гистограммы анализа нормального распределения для видов связи T1 - T4

Из Рис. 2 и Рис. 3 видно, что значения не одной из анализируемых выборок не находятся в пределах колоколообразной кривой, следовательно, анализируемые выборки не могут быть описаны нормальным распределением. Поэтому для исследования основных характеристик распределения анализируемой случайной величины адекватным является применение методов непараметрической статистики. Полученные результаты представлены в Табл. 5.

Таблица 5. Результат расчета основных статистических характеристик для Т1 - Т4

Описательные характеристики	Т1	Т2	Т3	Т4
Среднее значение выборки	2965924	9965559	19501476	42,14
Максимальное значение выборки	90341000	192513000	310000000	96,21
Минимальное значение выборки	0	18470	1160	0,05
Размах	90341000	192494530	309998840	96,16
Мода	–	–	–	–
Медиана	540850	2490022	5828157	40,41
Квартиль Q_1	108804	587642	1817835	23,55
Квартиль Q_2	557167	2303484	5952441	41,39
Квартиль Q_3	1830508	5396746	15414791	60,71
Среднее линейное отклонение	2886107	9799232	19350479	41,55

Из Таблицы 5 видно, что среднее количество точек подключения (ТП) для ФПШС составляет 2 965 924, при нулевом минимуме, с максимумом в 90341000 ТП. Среднее значение ФТС – 9 965 559 ТП, МС – 19 501 476 ТП.

Следовательно, в настоящее время наиболее широко распространенной в анализируемых странах является мобильная связь, среднее количество ТП для которой практически в два раза больше, чем у ФТС и в 10 раз превышает среднее значение пользователей ФПШС. Это можно пояснить тем, что широкополосная связь пока еще не достигла максимального развития, платежеспособный спрос на эти услуги еще не успевает за предложением.

Из Таблиц 1 - 4 видно, что наиболее часто встречающегося значения в анализируемой выборке не существует, поэтому моду определить невозможно, однако серединой выборки для Т1 является значение 540 850, для Т2 – 2 490 022, для Т3 – 5 828 157, отсюда видно, что использование фиксированной телефонной связи по-прежнему актуально и пока превосходит в 4 раза использование фиксированной проводной широкополосной связи.

Помимо этого в таблице представлены три значения, которые делят исследуемую выборку на четыре равные части (квартили Q_1, Q_2, Q_3) что, например, для ФТС Q_3 составляет 5 396 746 , а это означает, что значения количества ТП достаточно сконцентрированы относительно середины выборки.

Для того чтобы учесть различия значений исследуемой выборки рассчитано среднее линейное отклонение (СЛО), которое для ФПШС составило 2 886 107, для ФТС – 9 799 232, для МС – 19 350 479. Эта характеристика показывает величину разброса значений исследуемых случайных величин относительно среднего значения выборки, например, для Т1 СЛО составило 2 886 107 относительно середины выборки, это объясняется тем, что значения ТП для этого вида связи существенно отличаются друг от друга, поскольку этот вид связи еще не достиг максимального развития и его использование существенно отличается в различных странах, в соответствии с их экономическим развитием в целом.

На следующем этапе выполняется анализ признаков имеющейся выборки, для этого разработано множество методов АД, к которым относятся, в частности, факторный, корреляционный, кластерный и регрессионный анализы, которые можно использовать как независимо друг от друга, так и в совокупности.

Факторный анализ [Иберла, 1980] позволяет оценить влияние отдельных факторов на результирующие характеристики процесса и включает пять этапов. На первом этапе отбираются факторы, определяющие исследуемые показатели.

Следующим этапом является классификация и систематизация факторов с целью обеспечения комплексного системного подхода к исследованию их влияния на исследуемую случайную величину.

Далее определяется форма зависимости и моделируется взаимосвязь между факторами и результативными показателями. На четвертом этапе рассчитывается влияние факторов и оценка роли каждого из них в изменении величины результативного показателя. Заключительным этапом является практическая работа с факторной моделью и собственно факторный анализ.

Последовательность выполнения этих этапов отображает Рис. 4.

Рисунок 4. Последовательность выполнения факторного анализа

В качестве примера приведены результаты использования методов факторного анализа при выявлении степени влияния отдельных факторов на процесс развития инфокоммуникационных технологий.

В качестве исходных данных использована статистическая информация по макроэкономической и гендерной статистике для 53-х стран мира *UNECE* [UNECE Statistic Database]. Макроэкономическая информация – это совокупность экономических показателей для стран всего мира, а гендерная статистика представляет собой социальные данные, классифицированные по половому признаку. Информацию, полученную из баз данных *UNECE,* можно охарактеризовать параметрами, выбор которых обусловлен тем, что на сайте Европейская экономическая комиссия Организации Объединенных Наций (ЕЭК ООН) представлены их реальные значения, а также, тем что доступ к другим источникам аналогичной информации затруднителен.

Проанализированы следующие факторы:

1. Население страны: территория; общая численность населения; плотность населения, чел. на кв. км.

2. Численность населения страны в отдельных возрастных группах по полу: мужчины возрастных групп 0-14, 15-64, 64+ и женщины тех же возрастных групп, процентная составляющая мужчин и женщин в общем населении страны.

3. Доля населения страны, использующего персональный компьютер (ПК), разделенная по половой принадлежности и распределенная в отдельные группы по возрасту и полу: мужчины в возрасте 16-24, 25-54, 55-74 и женщины тех же возрастных групп.

4. Занятость населения, с точки зрения международной стандартной отраслевой классификации (МСОК) и безработица населения: уровень безработицы молодежи; занятость (МСОК версия 3.1 раздел A и B) *UNECE* [UNECE Publications]; занятость в промышленности и электроэнергии (МСОК версия 3.1 C-E); занятость (МСОК версия 3.1 разделы G-I и J-K); занятость в других услугах (МСОК версия3.1 L-P).

5. Обобщенные экономические показатели в международно-сопоставимых ценах, включающие:
   - Валовой внутренний продукт (ВВП) в текущих ценах и паритет покупательской способности (ППС) в млн. долл. США;
   - ВВП в текущих ценах, млн. национальной валюты;
   - ВВП на душу населения в ценах и ППС (США) текущего года, долларов США;
   - ВВП на душу населения в ценах текущего года, единиц национальной валюты.

6. Обобщенные экономические показатели, включающие: уровень безработицы; занятость и темпы ее роста; индекс потребительских цен и темпы его роста; паритет покупательной способности (ППС); расходы на конечное потребление инфокоммуникационных услуг (ИКУ) на душу населения.

7. Экономическая деятельность:
   - Доля экспорта и импорта товаров и услуг в процентах от ВВП;
   - Внешний баланс товаров и услуг, % от ВВП;
   - Внешний долг, млн. долл. США; ВВП: сельское хозяйство, промышленность (включая строительство), услуги: производственный метод, индекс 2005 = 100, ценах и ППС 2005;
   - ВВП: МСОК A-B, C-E, F, G-I, J-K, L-P, производственный метод, доля в процентах от валовой добавленной стоимости (ВДС), ценах и ППС текущий год.

Факторный анализ, независимо от используемых методов, начинается с определения количества факторов, которые имеет смысл использовать для анализа. Этот этап является очень важным, так как исследователь зачастую имеет большой набор характеристик, но заранее неизвестно, сколько факторов необходимо и достаточно для предоставления данного набора характеристик.

Сама же процедура факторного анализа предполагает предварительное задание числа факторов. Исходя из этого, исследователь должен заранее определить или оценить возможное количество факторов. Для определения числа факторов удобно использовать программный пакет для статистического анализа информации „*STATISTICA-8*", с помощью которого в нашем случае рассчитано два критерия: метод главных компонент и критерий каменой осыпи.

Одним из наиболее распространенных критериев поиска факторов является метод главных компонент. Оценка значимости главных компонент производится по их собственным значениям, представляющим собой дисперсию стандартизированных исходных данных. Основная идея метода основана на следующем предположении: чем выше % общей дисперсии, тем больше информации содержит значение. Поэтому метод главных компонент сводится к последовательной процедуре: вначале ищется первый фактор, объясняющий наибольшую часть дисперсии, затем независимый от него второй фактор, объясняющий наибольшую часть оставшейся дисперсии, и т.д.

Для работы с программным пакетом „*STATISTICA-8*" необходимо ввести все значения исследуемой случайной величины в главный модуль, после чего запускается модуль „Анализ главных компонент", после чего выбираются переменные для анализа (в нашем случае это параметры представленные выше). После того как переменные заданы, необходимо принять решение как будет проводится анализ: на основе ковариаций, либо корреляций. Ковариация несет тот же смысл, что и коэффициент корреляции – она показывает, есть ли линейная взаимосвязь между двумя случайными величинами, и может рассматриваться как „двумерная дисперсия". Знак ковариации указывает на вид линейной связи между рассматриваемыми величинами: если она больше 0 – это означает прямую связь (при росте одной величины растет и другая), ковариация <0 указывает на обратную связь. При ковариации равной 0 линейная связь между переменными отсутствует. В нашем случае анализ проведен на основе корреляционной матрицы, поэтому, выбрана опция „Анализ основан на корреляциях", а для того, чтобы заменить пропущенные значения анализируемых величин на их средние значения, в группе опций

„Удаление пропущенных данных” установлено значение „Замена средним”. После этого нажав кнопку „ОК” получаем результаты, представленные на Рисунке 5.

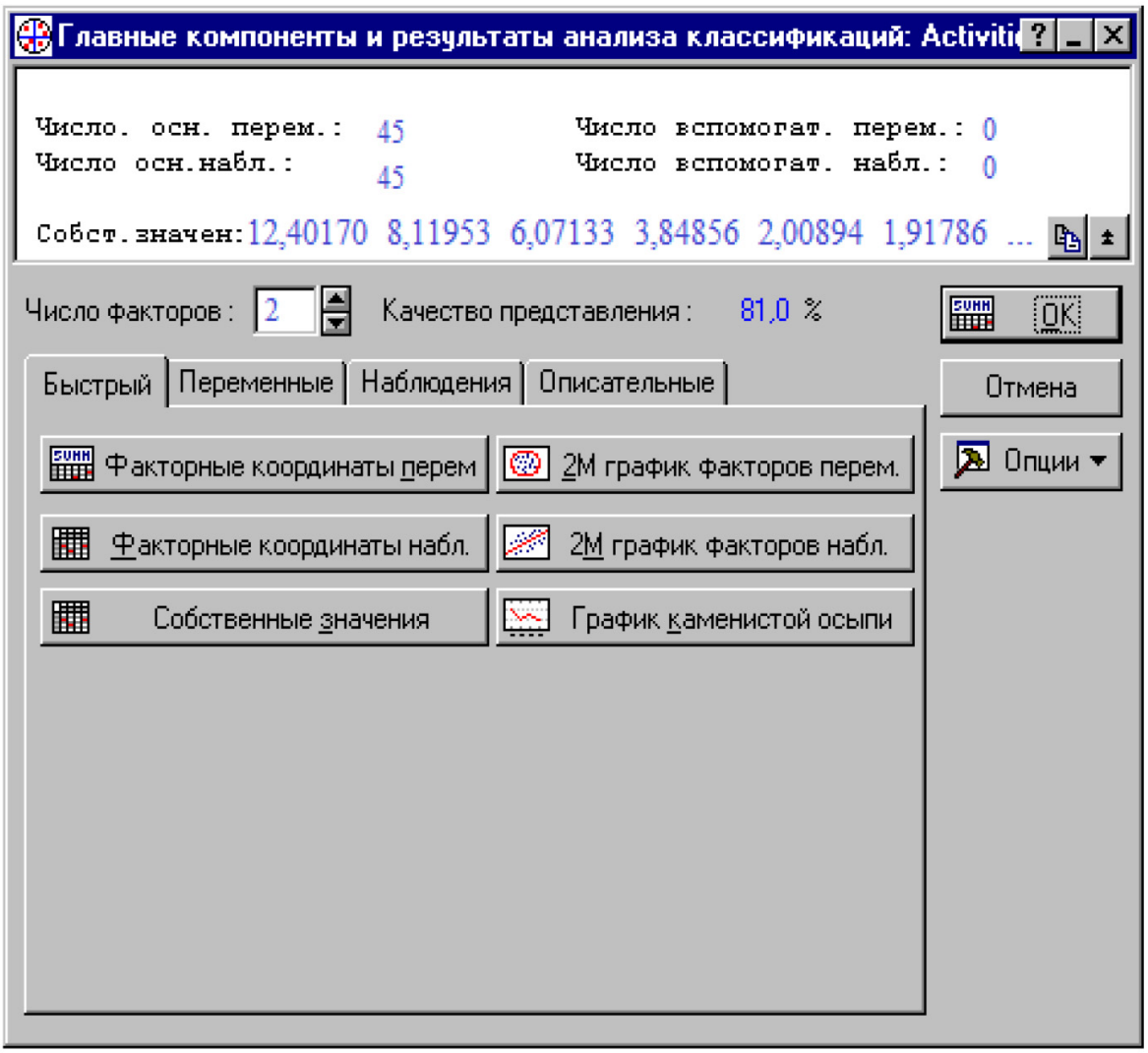


Рисунок 5. Форма модуля “Анализ главных компонент”

В информационном поле диалога „Главные компоненты и результаты анализа классификаций” представлена общая информация о результатах текущего анализа. Это число активных и вспомогательных переменных, наблюдений и собственные значения. При нажатии кнопки „Собственные значения”, получаем таблицу собственных значений (Табл. 6), в которой для каждого собственного значения представлен процент объясненной дисперсии, кумулятивное собственное значение и кумулятивный процент объясненной дисперсии.

Первое собственное значение всегда является наибольшим и превышает единицу; это определяется алгоритмом работы процедуры – факторы вычитаются в порядке их влияния на дисперсию переменных. затем вычисляется дисперсия, обуславливаемая определенным фактором. С изъятием каждого нового фактора собственные значения уменьшаются, а кумулятивный процент приближается к 100.

Таблица 6. Собственные значения матрицы корреляции

Фактор	Собственное значение параметров	% общей дисперсии	Кумулятивное собственное значение	Кумулятивный процент
1	12,40170	27,55932	12,40170	27,55932
2	8,11953	18,04340	20,52123	45,60273
3	6,07133	13,49185	26,59256	59,09457
4	3,84856	8,55236	30,44112	67,64694
5	2,00894	4,46431	32,45006	72,11125
6	1,91786	4,26191	34,36792	76,37316
7	1,61008	3,57796	35,97800	79,95112
8	1,46741	3,26090	37,44541	83,21202
9	1,17164	2,60364	38,61705	85,81566
10	1,10646	2,45880	39,72351	88,27446

Как видно из Таблицы 6, собственное значение для первого фактора равно 12,4, то есть этот фактор описывает примерно 27,6% общей выборки. Второй фактор, имеющий значение 8,12 отвечает за 18,04% общей выборки и т.д.

В рассматриваемом примере оставляем для дальнейшего исследования только первые пять факторов, которые имеют значение больше двух, в соответствии с критерием Кайзера [Kaiser, 1960].

Другим способом определения наиболее значимых факторов является построение и анализ графика каменной осыпи (Рис. 6), на котором отображается последовательность собственных значений факторов. Р. Б. Кеттел [Cattel, 1966] предложил определить на графике собственное значение, начиная с которого „горка" теряет свою кривизну и выходит на примерно постоянный уровень. Правая часть графика является лишь незначительными остатками „каменистой осыпи". Таким образом, при использовании такого подхода нужно оставить только факторы, расположенные слева от „каменной осыпи". На Рис. 6 показан график, на котором точка, где непрерывное падение собственных значений замедляется, может соответствовать пятому или шестому фактору.

В результате сопоставления таблицы собственных значений и графика „каменной осыпи" определено, что в анализируемом случае целесообразно выделить пять факторов.

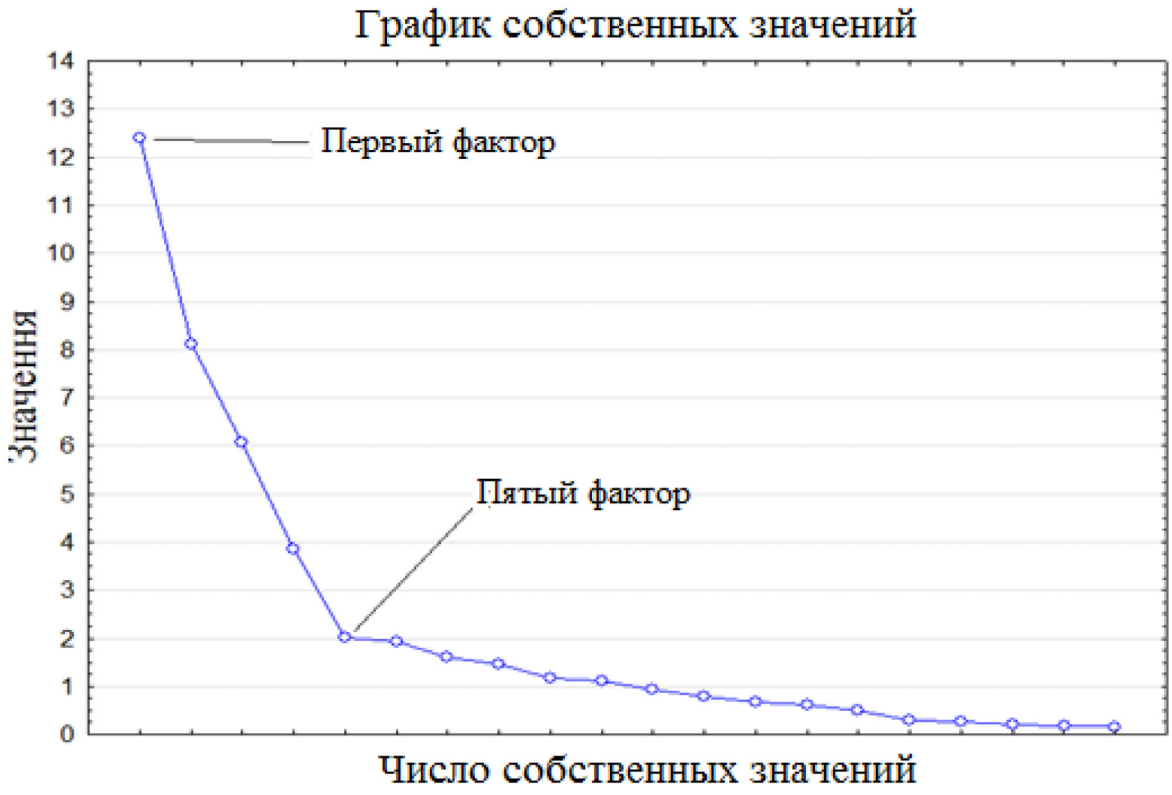


Рисунок 6. График собственных значений факторов

Основные результаты факторного анализа выражаются в наборах факторных нагрузок, то есть значений коэффициентов корреляции каждого из исходных признаков каждым из выявленных факторов и факторных рычагов. Факторными рычагами называют количественные значения выделенных факторов для каждого из *n* имеющихся объектов. Чем теснее связь конкретного признака с каким-то фактором, тем выше значение факторной нагрузки. Положительный знак факторной нагрузки указывает на прямую (а отрицательный знак – на обратную) связь данного признака с фактором.

В примере использовано несколько видов факторных нагрузок, касающиеся способов вращения, целью которых является получение интерпретируемой матрицы нагрузок, то есть факторов, отмеченных высокими нагрузками для некоторых переменных и низкими для других. При этом использованы методы без обращения, варимакс и квартимакс [Лоули, 1967]. Варимакс [Ким, 1989] – наиболее часто используемый на практике метод, цель которого минимизировать количество переменных, имеющих высокие нагрузки на данный фактор, что способствует упрощению описания фактора за счет группировки вокруг него только тех переменных, которые с ним связаны в большей степени, чем остальные.

Квартимакс [Ким, 1989] – в определенном смысле противоположен варимаксу, т.к. минимизирует количество факторов, необходимых для объяснения данной переменной. Поэтому он усиливает интероперабельность переменных. Квартимакс – вращение приводит к выделению одного из общих факторов с достаточно большими нагрузками на большинство переменных. Пример результата расчетов для одного из этих методов приведен в Таблице 7, где приняты следующие обозначения: М – мужчины, Ж – женщины, ВГ – возрастная группа.

При анализе Таблицы 7, можно заметить, что начальные параметры объединились, образуя факторы. Например, общая численность населения и возрастные группы образуют фактор два, а общая численность мужчин и женщин объединились в фактор три, и так далее. Исходя из результатов Таблицы 7, в качестве основных выделено следующие пять факторов: использование ПК в производственных целях, численность населения, производственная необходимость, экономический фактор и уровень безработицы.

Корреляционный анализ [Ферстер, 1983] двух случайных величин включает в себя: построение корреляционного поля и составление корреляционной таблицы, вычисление выборочных коэффициентов корреляции и корреляционных отношений, а также проверку статистической гипотезы значимости связи (Рис. 7).

Таблица 7. Факторы нагрузок без вращения

Переменные	Фактор 1	Фактор 2	Фактор 3	Фактор 4	Фактор 5
Общая численность населения	-0,60644	0,784758	-0,048107	0,028718	0,027946
ВГ 0-14, М	-0,56028	0,805252	-0,035146	0,023690	-0,008963
ВГ 0-14, Ж	-0,55984	0,805216	-0,034697	0,023430	-0,008984
ВГ 15-64, М	-0,60767	0,782602	-0,049392	0,028653	0,032558
ВГ 15-64, Ж	-0,60305	0,783924	-0,051367	0,024812	0,032395
ВГ 64+, М	-0,66333	0,718933	-0,046818	0,062626	0,061124
ВГ 64+, Ж	-0,64517	0,725005	-0,059302	0,033788	0,063937
Общее население, М (%)	-0,10863	-0,047701	-0,720109	0,002136	-0,107401
Общее население, Ж (%)	-0,08897	-0,006833	-0,762461	-0,033656	-0,016059
Использование ПК, 16-24, М	-0,71622	-0,432270	-0,212928	-0,051869	0,130498
Использование ПК, 16-24, Ж	-0,73283	-0,440734	-0,197314	-0,051307	0,141277
Использование ПК, 25-54, М	-0,79541	-0,431768	-0,174681	-0,113321	0,047776
Использование ПК, 25-54, Ж	-0,79837	-0,426325	-0,161282	-0,116773	0,061987

Переменные	Фактор 1	Фактор 2	Фактор 3	Фактор 4	Фактор 5
Использование ПК, 55-74, М	-0,80783	-0,358019	-0,093864	-0,195795	-0,094792
Использование ПК, 55-74, Ж	-0,78132	-0,322063	-0,082339	-0,190662	-0,077259
ВВП в текущих ценах, (млн. национальной валюты)	0,17189	0,309228	-0,089602	-0,834416	0,294933
ВВП на душу населения в ценах текущего года, (единиц нац. валюты)	0,18696	0,076216	-0,104545	-0,820006	0,308456
Расходы на конечное потребление ИКУ на душу населения, (дол. США)	-0,85094	-0,269061	-0,070655	-0,056333	-0,031010
ППС, единиц национальной валюты за долл. США	0,32237	0,217275	-0,085943	-0,804335	0,249007
Внешний долг, (млн. долл. США)	-0,71501	0,490130	-0,008225	0,064208	-0,003780
ВВП: сельское хозяйство	-0,11039	-0,080288	-0,802218	-0,117216	-0,164368
ВВП: промышленность	-0,02994	0,029975	-0,817944	-0,138809	-0,257610
ВВП услуги	-0,01826	0,039206	-0,840393	-0,105549	-0,267240
ВВП: МСОК A-B	0,70899	0,321233	-0,323788	0,058450	-0,191523
ВВП: МСОК J-K	-0,76247	-0,299351	-0,202627	0,030646	0,129242
Занятость (МСОК вер. 3.1 J-K)	-0,77092	-0,345382	0,331043	-0,011503	-0,049646

Рисунок 7. Последовательность корреляционного анализа

Основное назначение корреляционного анализа – выявление связи между двумя или более изучаемыми переменными, которая рассматривается как совместное согласование изменения двух исследуемых характеристик. Такая изменчивость описывается тремя основными характеристиками: формой, направлением и силой. По форме корреляционная связь может быть линейной или нелинейной.

Более удобной для выявления и интерпретации корреляционной связи является линейная форма. Для линейной корреляционной связи выделяют два основных направления: положительное („прямая связь”) и отрицательное („обратная связь”). Сила связи указывает, насколько ярко проявляется общая изменчивость исследуемых переменных. В качестве числовой характеристики вероятностной связи используют коэффициенты корреляции, значения которых изменяются в диапазоне от -1 до +1 [Ферстер, 1983]

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \quad (1)$$

где x_i – значение, принимаемое в выборке *X*;

y_i – значение, принимаемое в выборке *Y*;

$\bar{x}$ и $\bar{y}$ – средние значение выборок *X* и *Y*.

После проведения расчетов, как правило, отбираются только самые сильные корреляции, которые затем интерпретируются. Критерием для отбора „достаточно сильных" корреляций может быть, как абсолютное значение самого коэффициента корреляции (от 0,6 до 1), так и относительная величина этого коэффициента, определяемого по уровню статистической значимости (от 0,01 до 0,1), зависящая от размера выборки [Ферстер, 1983].

Таким образом, задача корреляционного анализа сводится к установлению направления и формы связи между варьируемыми признаками, измерению ее тесноты и к проверке уровня значимости полученных коэффициентов корреляции. Используется множество коэффициентов корреляции, чаще всего применяют коэффициенты *r*-Пирсона, *r*-Спирмена и *τ*-Кендалла. Выбор метода вычисления коэффициента корреляции зависит от типа шкалы, к которой относятся переменные (Табл. 8). Для переменных с интервальной и номинальной шкалой используется коэффициент корреляции Пирсона, а если, по меньшей мере, одна из двух переменных имеет порядковую шкалу или не может быть описана нормальным распределением, используется ранговая корреляция по Спирмену или Кендаллу.

Для условий примера, рассмотренного выше, проанализируем применение методов корреляционного анализа с целью выявления зависимостей между найденными ранее факторами и рассматриваемыми видами связи. В ходе подготовки данных к проведению расчетов на первом этапе выявлено отсутствие нормального распределения, в связи с этим используются методы непараметрической статистики, результаты расчета основных характеристик которых представлены в Табл. 9 в виде матрицы.

Таблица 8. Меры расстояний

Типы шкал		Мера связи
Переменная	Переменная	
Интервальная или отношений	Интервальная или отношений	Коэффициент Пирсона
Ранговая, интервальная или отношений	Ранговая, интервальная или отношений	Коэффициент Спирмена
Ранговая	Ранговая	Коэффициент Кендалла
Дихотомическая	Дихотомическая	Коэффициент φ, четырехполевая корреляция
Дихотомическая	Ранговая	Рангово-бисериальный коэффициент
Дихотомическая	Интервальная или отношений	Бисериальный коэффициент
Интервальная	Ранговая	Не разработаны

Таблица 9. Таблица расчета непараметрической статистики

	AS	D	MSD	MaxS	MinS	R	M
Фактор 1	0,518	0,114	0,338	1	0	1	0,620
Фактор 2	0,013	0,001	0,026	0,167	0	0,167	0,004
Фактор 3	0,776	0,054	0,232	1	0	1	0,843
Фактор 4	0,042	0,022	0,147	1	0,00092	0,999	0,005
Фактор 5	0,315	0,044	0,209	1	0,02021	0,979	0,274
T1	0,044	0,015	0,121	0,831	0	0,831	0,008
T2	0,034	0,008	0,087	0,588	0,00007	0,588	0,007
T3	0,037	0,005	0,073	0,419	0,00005	0,419	0,012
T4	0,500	0,072	0,269	1,000	0,02270	0,977	0,492

В Табл. 9 использованы следующие показатели:

- *AS* (*average of sample*) – среднее значение выборки,
- *D (dispersion)* – дисперсия,
- *MSD (mean square deviation)* – среднеквадратическое отклонение,
- *MaxS (maximum sample)* – максимальное значение,
- *MinS (minimum sample)* – минимальное значение,
- *R (range)* – размах,
- *M (median)* – медиана.

Результирующие значения линейной корреляции представлены в виде Табл. 10, где каждая ячейка содержит значение коэффициента корреляции двух выборок.

При значении коэффициента корреляции больше или равно 0,6 [Ферстер, 1983], можно утверждать, что между двумя массивами, которые исследуются, существует зависимость. В Табл. 10 соответствующие значения, указывающие на наличие линейной корреляции, выделены жирным шрифтом.

По результатам рассчитанных значений корреляции, выделено следующие зависимости:

1. Очень высокое значение расчетного показателя корреляции, практически равное единице, отмечено между ФПШС (Т1) и ФТС (Т2).
2. Средний показатель корреляции присутствует между фактором первым и использованием Интернет. Это обусловлено тем, что большинство компьютеров, используемых в производственных целях, подключено к Интернет.
3. Достаточно высокий показатель корреляции имеют фиксированная проводная широкополосная связь, фиксированная телефонная связь и мобильная связь. Данный показатель объясняется тем, что количество пользователей любой технологии имеет прямую зависимость от количества населения.
4. Исходя из описанного выше и прямой зависимости количества пользователей ФПШС и ФТС высокое значение коэффициента корреляции между Т3 и Т1, Т2 обусловлено потребностью населения в связи независимо от их местоположения.

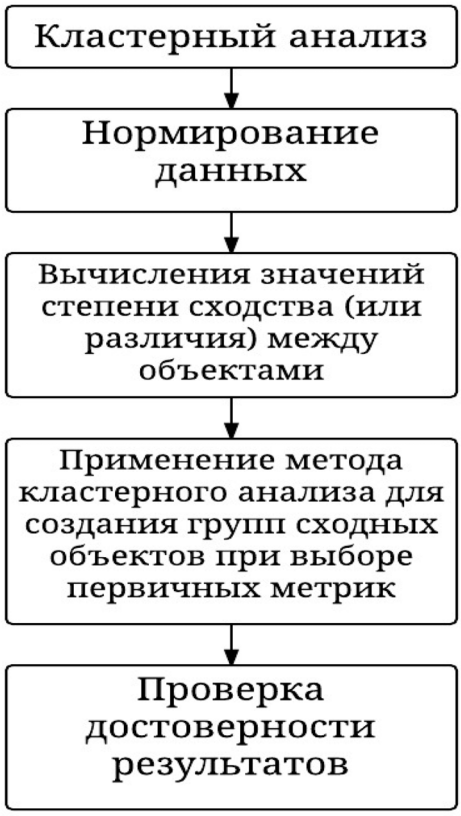


Рисунок 8. Последовательность кластерного анализа

Таблица 10. Таблица рассчитанных значений корреляции

	Фактор 1	**Фактор 2**	**Фактор 3**	**Фактор 4**	**Фактор 5**	**T1**	**T2**	**T3**	**T4**
Фактор 1	1,00	0,17	0,18	-0,28	-0,12	0,21	0,20	0,23	***0,65***
Фактор 2	0,17	1,00	0,06	-0,05	-0,11	***0,93***	***0,97***	***0,98***	0,11
Фактор 3	0,18	0,06	1,00	0,10	0,25	0,00	0,02	0,09	-0,14
Фактор 4	-0,28	-0,05	0,10	1,00	-0,20	-0,08	-0,07	-0,08	-0,22
Фактор 5	-0,12	-0,11	0,25	-0,20	1,00	-0,11	-0,10	-0,10	-0,18
T1	0,21	***0,93***	0,00	-0,08	-0,11	1,00	***0,99***	***0,89***	0,28
T2	0,20	***0,97***	0,02	-0,07	-0,10	***0,99***	1,00	***0,94***	0,22
T3	0,23	***0,98***	0,09	-0,08	-0,10	***0,89***	***0,94***	1,00	0,15
T4	***0,65***	0,11	-0,14	-0,22	-0,18	0,28	0,22	0,15	1,00

Кластерный анализ представляет собой совокупность математических методов, предназначенных для формирования отдельных групп объектов похожих по информации, свойствами или другим критериям [Дюран, 1977], такие группы объектов называют кластерами. Кластерный анализ широко используется в науке как средство типологического анализа, фактически кластерный анализ – это обобщенное название большого набора алгоритмов, используемых при классификации объектов, Рис. 8.

В любой научной деятельности классификация является одной из фундаментальных составляющих, без которой невозможны построение и проверка научных теорий и гипотез.

Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма [Tryon, 1939].

Для применения кластерного анализа необходимо проверить удовлетворяют данные следующим требованиям:

1. Выборка должна быть однородной.
2. Показатели не должны быть связаны между собой и не должны быть безразмерными.
3. Распределение показателей должно быть близким к нормальному.
4. Отсутствие влияния на значение показателей случайных факторов.

После получения результатов и их последующего анализа возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата, определяемого высокой степенью сходства объектов внутри каждого кластера. Результатом применения процедуры кластеризации [Мандель, 1988] может быть формирование нескольких подгрупп кластеров объектов исследования, в каждом из которых содержится объекты наблюдения „похожи" по некоторым выбранным показателям или по интегрированному показателю.

Для определения схожести используют следующие метрики:

1. Эвклидово расстояние [Мандель, 1988] – наиболее общий тип расстояния, являющийся геометрическим расстоянием в многомерном пространстве

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2} \quad (2)$$

2. Квадрат эвклидова расстояния [Мандель, 1988] – используется для придания большего веса более отдаленным друг от друга объектам

$$\rho(x, x') = \sum_{i}^{n} (x_i - x_i')^2 \quad (3)$$

3. Расстояние городских кварталов (манхэттенское расстояние) [Мандель, 1988] – это просто среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для расстояния Евклида. Однако для этой меры влияние отдельных выбросов уменьшается, так как они не возводятся в квадрат

$$\rho(x, x') = \sum_{i}^{n} \left| x_i - x_i' \right| \tag{4}$$

4. Расстояние Чебышева [Мандель, 1988] полезно использовать, когда необходимо определить два объекта как „различные", если они отличаются по какой-либо одной координате

$$\rho(x, x') = \max(\left| x_i - x_i' \right|) \tag{5}$$

5. Степенное расстояние [Мандель, 1988] используют в том случае, когда необходимо прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются

$$\rho(x, x') = \sqrt{\sum_{i}^{n} (x_i - x_i')^p} \tag{6}$$

В примере применения кластерного анализа использован интегрированный показатель, в качестве которого выбран средний показатель ТП для каждого из видов связи. Для выявления групп стран, близких по уровню принятия интегрированного показателя, применен метод „ближнего соседа" при использовании трех видов метрик: Евклидовой, Чебышева и Манхэттенской.

В Таблице 11 рассмотрены начальные статистические данные для каждой группы показателей, и видно, что они имеют большой разброс, что негативно влияет на построение кластера. Для уменьшения разброса начальных данных применяется процедуру их нормирования.

Таблица 11. Данные о количестве пользователей

Страны	Фиксированная проводная широкополосная связь	Мобильная телефонная связь	Фиксированная телефонная связь	Процент пользователей Интернет
1	2	3	4	5
Австрия	1262746	9509198	3614769	60
Азербайджан	267279	4476405	1208219	21
…	…	…	…	…
Молдова	119321	1718217	955713	19
Польша	2690346	31933336	10111327	41
США	53646550	218217541	167908456	67
1	2	3	4	5
Туркменистан	753	1132026	444938	2
Узбекистан	51350	8125834	1801042	11
Украина	1132626	33905591	12011422	11
Франция	12018732	49366166	36033248	53
Швейцария	1791004	7615364	4989000	72

Полученные значения расстояний между кластерами стран с использованием Евклидовой метрики представлены в Табл. 12, при этом учтена симметричность Евклидовой метрики, согласно которой расстояние, например, между Молдовой и Австрией равна расстоянию между Австрией и Молдовой, поэтому Евклидово расстояние рассчитано только для одной из каждой пары стран.

Таблица 12. Матрица расстояний между странами

Страны	**Австрия**	**…**	**Молдова**	**Россия**	**США**	**Узбекистан**	**Украина**	**Франция**	**Швейцария**
Австрия	0	…	0,50	0,74	1,66	0,60	0,60	0,33	0,15
...	…	**0**	…	…	…	…	…	…	…
Молдова	0,50	…	0	0,67	1,81	0,10	0,19	0,55	0,64
Россия	0,74	…	0,67	0	1,32	0,66	0,53	0,54	0,86
США	1,66	…	1,81	1,32	0	1,83	1,72	1,36	1,68
Узбекистан	0,60	…	0,10	0,66	1,83	0	0,14	0,62	0,74
Украина	0,60	…	0,19	0,53	1,72	0,14	0	0,56	0,75
Франция	0,33	…	0,55	0,54	1,36	0,62	0,56	0	
Швейцария	0,15	…	0,64	0,86	1,68	0,74	0,75	0,40	0

После получения значений Евклидовых расстояний между странами проведена кластеризация объектов с помощью программного пакета статической обработки исходных данных „*STATISTICA-8*", в результате которой выявлены группы стран, обобщенный уровень которых, по признакам Т1 - Т4, достаточно близок. Для большей наглядности приведены дендриты, иллюстрирующие содержание кластеров (Рис. 9 и Табл. 11, 12), описывающие объединение стран в кластеры 1, 2 и 3 по различным показателям.

Таблица 13. Страны, составляющие кластер № 1

Страна	Среднее количество пользователей / коэффициент нормирования данных за 10 лет для:							
	Т1	**Коэф.**	**Т2**	**Коэф.**	**Т3**	**Коэф.**	**Т4**	**Коэф.**
Кластер 1								
Франция	12019886	0,22	36090709	0,22	49366166	0,23	52	0,63
Турция	3434316	0,06	17657764	0,11	46403659	0,21	24	0,29
Великобритания	11225663	0,21	34046706	0,20	66470794	0,30	68	0,82
Германия	14708381	0,27	52743846	0,31	84013461	0,39	65,83	0,79
США	53849396	1	167861456	1	218217541	1	66	0,79
Россия	6647832	0,12	40524691	0,24	135816791	0,60	23,06	0,28
Испания	6062649	0,11	19073325	0,11	42725493	0,20	47,28	0,57
Италия	7469673	0,14	24586898	0,15	75095222	0,34	39,36	0,48

Кластер 1 представляет собой группу стран, имеющих наибольшее значение показателей T1 - T4. Эти страны представлены в Табл. 13. Кластер 2 представляет собой группу стран со средним уровнем развития видов связи. В этом кластере, уровень развития примерно в 1, 2 раза ниже, чем в кластере 1. Данные предоставлены в Табл. 14.

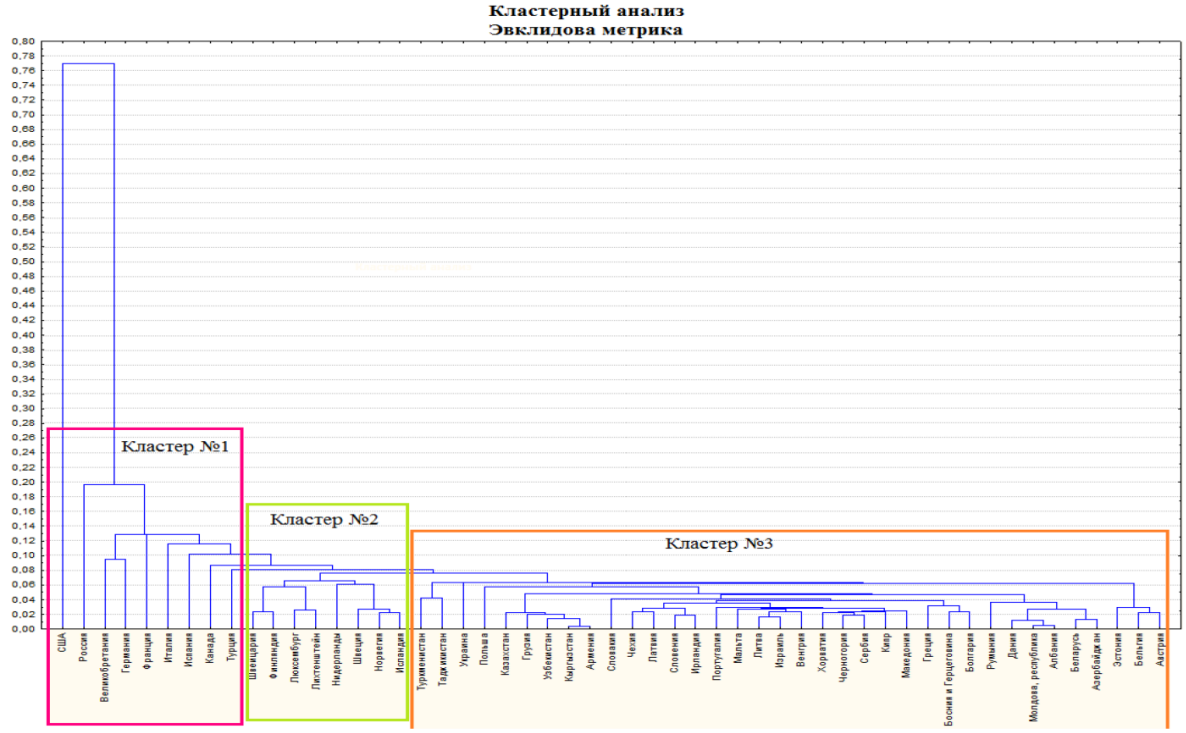


Рисунок 9. Кластер №1, 2, 3

В кластер 3 включены страны, в которых уровень развития ФПШС, МС, ФТС и доступа к Интернет в три раза ниже, чем в кластере 1. Список стран можно определить из общей выборки, исключив из нее страны, принадлежащие к кластерам 1 и 2.

Таблица 14. Страны, составляющие кластер № 2

Страна	Среднее количество пользователей / коэффициент нормирования данных за 10 лет для:							
	T1	**Коэф.**	**T2**	**Коэф.**	**T3**	**Коэф.**	**T4**	**Коэф.**
Кластер 2								
Исландия	71939	0,001	191338	0,001	300933	0,001	82,7	1,000
Лихтенштейн	10273	0,0002	19707	0,0001	26973	0,0001	65,9	0,796
Люксембург	89519	0,002	255968	0,0015	598625	0,003	67,9	0,821
Нидерланды	4100413	0,076	7719577	0,046	16530562	0,0758	76,2	0,921
Канада	7265525	0,11	18537085	0,04	19230156	0,07	71,4	0,86
Польша	2658205	0,049	10097934	0,06	31933336	0,15	40,5	0,489
Украина	1338558	0,025	1201422	0,07	33905591	0,16	10,9	0,131
Норвегия	1037630	0,019	1989249	0,012	4726491	0,022	76,2	0,921
Финляндия	1064900	0,020	195275	0,012	6181239	0,028	73,2	0,885
Швейцария	1788976	0,033	5090043	0,030	7615364	0,035	71,9	0,869
Швеция	2069195	0,038	5374828	0,320	9419770	0,043	80,2	0,969

Таким образом, по результатам кластерного анализа, можно сделать вывод о существовании трех основных групп стран с равным развитием четырех видов связи.

Первую группу составляют страны с высокоразвитым уровнем экономики. Это Великобритания, США, Италия, Турция, Россия, Франция, Канада, Испания, Германия.

Второй кластер объединяет высокоразвитые страны Центральной и Северной Европы: Исландию, Лихтенштейн, Люксембург, Нидерланды, Норвегию, Финляндию, Швейцарию, Швецию.

Наиболее многочисленным является кластер 3, объединяющий страны, которые быстро развиваются, это Австрия, Азербайджан, Албания, Армения, Беларусь, Бельгия, Болгария, Босния и Герцеговина, Македония, Венгрия, Греция, Грузия, Дания, Израиль Ирландия, Казахстан, Кипр, Кыргызстан, Латвия, Литва, Мальта, Молдова, Польша, Португалия, Румыния, Сербия, Словакия, Словения, Таджикистан, Туркменистан, Узбекистан, Украина, Хорватия, Черногория, Чехия, Эстония.

Результаты кластерного анализа позволяют также оценить степень различия между объектами и степень информационного неравенства между отдельными государствами. Государства, вошедшие в один кластер, имеют минимальный цифровой разрыв, под которым подразумевается разница между объемом использования сферы инфокоммуникаций в развитых и развивающихся странах.

Цифровой разрыв (*digital divide*) – понятие, получившее в последнее время широкое распространение в связи с возросшим значением новых информационно-коммуникационных технологий, усилением процессов глобализации, становлением информационного общества и переходом к глобальной экономике. Чем больше разница в номерах кластеров, в которые вошли отдельные государства, тем больше цифровой разрыв между ними.

Заключение

Ввиду интенсивного развития современного общества невозможно переоценить роль информации являющейся активным звеном всех сфер жизнедеятельности человека, которую целесообразно оценивать путем анализа имеющейся статистической информации. Однако в связи с большим объемом данных, большим количеством программ для обработки статистической информации и специалистов, плохо представляющих условия применения тех или иных методов анализа, оценку имеющейся информации выполнить не так просто.

Исследование существующих методов анализа статистических данных, учет условий применения каждого отдельного метода анализа, формирование методики и алгоритма анализа статистических данных позволяет корректно использовать методы анализа статистической информации для решения задач, относящихся к анализу функционирования и планирования развития информационных сетей.

Благодарности

Настоящая работа выполнена при поддержке интернационального проекта ITHEA XXI Института информационных теорий и их приложений FOI ITHEA и Ассоциации ADUIS Украина (Ассоциация разработчиков и пользователей интеллектуальных систем).

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (*www.ithea.org*) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (*www.aduis.com.ua*).

Литература

[Cattel, 1966] R.B. Cattel, “The scree test for the number of factors”, Multivariate Behavioral Research, 1966, №1, pp. 245-276

[Gannitskiy, 2010] Illiya Gannitskiy, “Adaptation time-series analysis to number of calls on the modern telecommunication network”, Modern Problems of Radio Engineering, Telecommunications and Computer Science: X int. conf.: proceedings Lviv: Publishing House of Lviv Polytechnic, 2010, 197 p.

[Gannytskyi, 2012] Illia Gannytskyi, “Short-term Forecasts Parameters the Stream of Calls on the Telecommunication Networks”, Modern problems of radio engineering, telecommunications and computer science: XIth int. conf.: Lviv: Publishing House of Lviv Polytechnic, 2012, 261 p.

[ITU] International Telecommunication Union ITU [Электронный ресурс], Режим доступа: http://www.ibm.com/. Дата обращения: 15.03.2015.

[Kaiser, 1960] H.F. Kaiser, “The application of electronic computers to factor analysis”, Educational and Psychological Measurement, 1960, № 20, pp.141 - 151

[Tryon, 1939] R.C.Tryon, “Cluster analysis”, London: Ann Arbor Edwards Bros, 1939, 139 p.

[UNECE Publications] United National Economic Commission for Europe UNECE Publications [Электронный ресурс], Режим доступа http://www.unece.org/publications, Дата обращения: 31.05.2015.

[UNECE Statistic Database] United National Economic Commission for Europe UNECE Statistic Database [Электронный ресурс], Режим доступа http://w3.unece.org/pxweb/, Дата обращения: 15.03.2015.

[Кендалл, 1976] Кендалл М., Стьюарт А. “Многомерный статистический анализ и временные ряды“, Москва: Наука, 1976, 375 с.

[Айвазян, 1989] Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д., “Прикладная статистика: Классификация и снижение размерности”, Москва: Финансы и статистика, 1989, 607 с.

[Айвазян, 1998] Айвазян С.А., Мхитарян В.С., “Прикладная статистика и основы эконометрики”, Москва: ЮНИТИ, 1998, 1004 с.

[Андерсон, 1963] Андерсон Т., “Введение в многомерный статистический анализ”, М.: Государственное издательство физико-математической литературы, 1963, 499 с.

[Андерсон, 1976] Андерсон Т., “Статистический анализ временных рядов”, М.: Мир, 1976, 757 с.

[Бондаренко, 2010a] Бондаренко А.А., “Оптимізація методів обробки результатів вимірювань”, Збірник тез IV Міжнародної НТК „Проблеми телекомунікацій – 2010”, Київ: КПІ., 2010, с. 246

[Бондаренко, 2010b] Бондаренко А.А., “Методи зберігання і обробки великих обсягів результатів вимірювань”, Збірник тез X Всеукраїнської НТК студентів і аспірантів „Інформаційні системи і технології”, Одеса, 2010, с. 59 - 60

[Бондаренко, 2011] Бондаренко А.А., “Некоторые аспекты обработки сверхбольших объемов статистических данных”, XI Всеукраїнська НТК студентів і аспирантів „Стан, досягнення і перспективи інформаційних систем і технологій”, Одеса: ОДАХ, 2011, с. 81 - 82

[Боровков, 2010] Боровков А.А., “Математична статистика”, М.: Лань, 2010, 704 с.

[Вайсфельд, 2006] Лободзинская И.Г., Вайсфельд Н.Д., Реут О.В., “Математична статистика”, Основи аналізу данних Навчально-методичний посібник, Одеса, 2006, 124 с.

[Гайворонская, 2007a] Гайворонская Г.С., Котелевец А.И., “Анализ использования беспроводных технологий в сетях абонентского доступа”, Труды III Международной научно-технической конференции „Сучасні інформаційно-комунікаційні технології” /COMINFO’ 2007, 24-28 вересня 2007, р.м. Ялта, смт. Лівадія

[Гайворонская, 2007b] Гайворонская Г.С., Ганницкий И.В., “Анализ параметров потоков вызовов на современной телекоммуникационной сети”, Труды III Международной научно-технической конференции „Сучасні інформаційно-комунікаційні технології” /COMINFO’ 2007, 24-28 вересня 2007, р.м. Ялта, смт. Лівадія

[Гайворонская, 2007c] Г. С. Гайворонская, И. В. Ганницкий, “Оценка влияния вероятностно-временной структуры потоков вызовов на объем оборудования телекоммуникационных сетей”, Проблеми телекомунікацій: перша наук.-техн. конф. : зб. тез., К. : НТУУ „КПІ”, 2007, С. 170 − 171

[Гайворонская, 2009a] Гайворонская Г. С., Котова А. И., “Исследование структурных характеристик абонентских линий непараметрическими методами обработки статистики”, Сборник тезисов IX Международной НТК „Математическое моделирование и информационные технологии”, Одесса: ОГАХ, 2009, С.102

[Гайворонская, 2009b] Г. С. Гайворонская, И. В. Ганницкий, “Анализ параметров потоков вызовов на телекоммуникационной сети”, Зв'язок., 2009, № 3 (87), С. 14 − 17.

[Гайворонская, 2009c] Г. С. Гайворонская, И. В. Ганницкий, “Обработка исходных данных при анализе параметров потоков вызовов на телекоммуникационной сети”, Холодильна техніка і технологія, Одеса: ОДАХ, 2009, № 3 (119), С. 73 − 76

[Ганницкий, 2008a] И. В. Ганницкий, “Особенности обработки статистических данных при анализе параметров потоков вызовов на современной телекоммуникационной сети”, Сучасні інформаційно-комунікаційні технології: IV міжнар. наук.-техн. конф.: зб. тез., К.: ДУІКТ, 2008, С. 86 - 87

[Ганницкий, 2008b] И.В. Ганницкий, “Влияние объема выборки на установление закона распределения длительности обслуживания потока вызовов на телекоммуникационной сети”, Математическое моделирование и информационные технологии: IX НТК.: сбор. тез., Одесса: ОГАХ, 2009, С. 104

[Ганницкий, 2009] И. В. Ганницкий, “Применение метода нормированного размаха для оценки характеристик потока вызовов на телекоммуникационной сети”, Цифрові технології, Одеса: ОНАЗ, 2009, Вип. 6, С. 71 − 76

[Ганницкий, 2010a] И. В. Ганницкий, “Обработка результатов измерений потоков вызовов в современной телекоммуникационной сети”, Сучасні інформаційно-комунікаційні технології: VI міжнар. наук.-техн. конф.: зб. тез., К. : ДУІКТ, 2010, С. 64 − 65

[Ганницький, 2010b] І. В. Ганницький, А. А. Бондаренко, “Розробка програмного забезпечення для підвищення швидкості обробки даних надвеликого об’єму”, Проблеми телекомунікацій: п’ята наук.-техн. конф.: зб. тез., К.: НТУУ „КПІ”, 2011, С. 107.

[Ганницький, 2011] І. В. Ганницький, А.А. Бондаренко, “Підвищення швидкості обробки результатів вимірювань параметрів потоку викликів”, Холодильна техніка і технологія, Одеса: ОДАХ, 2011, № 1 (129), С. 69 − 72

[Горяинов, 2001] В. Б. Горяинов, И. В. Павлов, Г. М. Цветкова, А. И. Тескин; Под ред. В.С. Зарубина, А.П. Крищенко, “Математическая статистика: Учебник для вузов”, М.: Издательство МГТУ им. Н.Э. Баумана, 2001, 424 с.

[Дюран, 1977] Дюран Б., Оделл П., “Кластерный анализ”, М.: Статистика, 1977, 128 с.

[Иберла, 1980] Иберла К., “Факторный анализ”, М.: Статистика, 1980, 398 с.

[Ким, 1989] Ким Дж.-О., “Факторный, дискриминантный и кластерный анализ”, М.: Финансы и статистика, 1989, 215 с.

[Кобзар, 2006] Кобзар А.И., “Прикладна математична статистика”, М.: Физматлит, 2006, 816 с.

[Котова, 2008] Г.С. Гайворонская, А. И. Котова, “Оценка параметров абонентских линий”, Збірник тез науково-технічної конференції „Сучасні інформаційно-комунікаційні технології”, К.: ДУІКТ, 2008, С. 62.

[Котова, 2009a] Г. С. Гайворонская, А. И. Котова, “Структурные характеристики абонентских линий”, Холодильна техніка і технологія, Одеса: ВЦ ОДАХ, 2009, № 4 (120), С. 78 – 81

[Котова, 2009b] Г.С. Гайворонская, А.И. Котова, “Исследование структурных характеристик абонентских линий непараметрическими методами обработки статистики”, Збірник тез IX міжнародної НТК „Математичне моделювання і інформаційні технології”, Одеса: ОДАХ, 2009, С. 102.

[Котова, 2010a] О.І. Котова, “Дослідження груп абонентських ліній сільських районів України”, Збірник наукових праць, К.: ВІТІ НТУУ „КПІ”, 2010, № 1, С. 11 – 15.

[Котова, 2010b] Г. С. Гайворонская, А. И. Котова, “Выбор технологии доступа на основании анализа структурных характеристик существующих абонентских сетей”, Холодильна техніка і технологія, Одеса: ВЦ ОДАХ, 2010, № 2 (122), С. 63 – 67.

[Леман, 1964] Леман Е., “Проверка статистических гипотез”, М.: Наука, 1964, 501 с.

[Ллойд, 1989] Ллойд Э., Ледерман В., “Справочник по прикладной статистике”, Том 1, М: Финансы и статистика, 1989, 510 с.

[Ллойд, 1990] Ллойд Э., Ледерман В., “Справочник по прикладной статистике”, Том 2, М: Финансы и статистика, 1990, 526 с.

[Лоули, 1967] Лоули Д., “Факторный анализ как статистический метод”, М. Мир, 1967, 144 с.

[Мандель, 1988] Мандель И.Д, “Кластерный анализ”, М: Финансы и статистика, 1988, 176 с.

[Окунь, 1974] Окунь Я., “Факторный анализ”, М. Статистика, 1974, 250 с.

[Орлов, 2004] Орлов А.И., “Прикладная статистика”, М.: Экзамен, 2004, 672 с.

[Павлов, 2007] Павлов С.В., “Снижение размерности параметров предоставления инокоммуникационных услуг методом факторного анализа”, Труды III Международной научно-технической конференции „Сучасні інформаційно-комунікаційні технології” COMINFO’ 2007, 24-28 вересня 2007, р.м. Ялта, смт. Лівадія

[Сахарова, 2008] С.В. Сахарова, “Исследование параметров сетей абонентского доступа”, Материалы VIII МНТК „Математическое моделирование и информационные технологии” ММИТ-2008, Одесса: ОГАХ, 2008, с.29.

[Ферстер, 1983] Ферстер Э., Ренц Б., “Методы корреляционного и регрессионного анализа”, Руководство для экономистов, М.: Финансы и статистика, 1983, 304 с.

[Харман, 1972] Харман Г., “Современный факторный анализ”, М.: "Статистика", 1972, 486 с.

Информация об авторах

Галина Сергеевна Гайворонская *– Институт холода, криотехнологий и экоэнергетики им. В. С. Мартыновского ОНАПТ, факультет информационных технологий и кибербезопасности, д.т.н., профессор, зав. кафедрой информационно-коммуникационных технологий, советник ректора по инфокоммуникациям; Украина, Одесса, 65026, ул. Дворянская, 1/3; тел. (048)-720-91-48;*

e-mail: gsgayvoronska@gmail.com

Основные направления научных исследований: оптимизация переходных перио-дов при эволюции телекоммуникационных сетей. Потоки вызовов, нагрузка и межузловое тяготение в сетях. Проблемы создания сетей доступа. Проблема построения полностью оптических сетей и систем коммутации.

Петр Петрович Яцук *– Национальная комиссия, осуществляющая государственное регулирование в сфере связи и информатизации, г. Киев, Украина;*

e-mail: petr_yatsuk@icloud.com

Основные направления научных исследований: телекоммуникационные сети и технологии, процессы реализации технологий передачи данных в телекоммуникационных сетях.

Юлия Сергеевна Казак *– Институт холода, криотехнологий и экоэнергетики им. В. С. Мартыновского ОНАПТ, факультет информационных технологий и кибербезопасности, аспирант кафедры информационно-коммуникационных технологий, ул. Дворянская, 1/3, г. Одесса, Украина, 65082, e-mail: flygerl1@bigmir.net*

Основные направления научных исследований: использование методов математической статистики в современных телекоммуникационных сетях.

Peculiarities Analysis of Statistical Information in ICT

Galina Gayvoronska, Petro Yatsuk, Yulya Kazak

Abstract: *Huge amounts of statistical information, which has become available at the information society creation, have led to the increase of number of this information analysis methods' incorrect usage. This often entails erroneous conclusions that may lead to negative consequences. Some guidance on the conditions of data analysis methods' usage are given in the work in order to avoid such a situation as well as electronic communications statistics' stepwise analysis method is formed.*

Keywords: *statistical information, information and communication technology, information and communication services.*

TABLE OF CONTENTS