**УДК** 311.2 + 616-036.22 + 004.8

А.В. Суворова, А.В. Лавренов,
Т.В. Тулупьева, А.Л. Тулупьев, А.Е. Пащенко

# МОДЕЛИРОВАНИЕ СОЦИАЛЬНО-ЗНАЧИМОГО ПОВЕДЕНИЯ РЕСПОНДЕНТОВ: АНАЛИТИЧЕСКАЯ И ЧИСЛЕННАЯ ОЦЕНКИ ИНТЕНСИВНОСТИ В ОКРЕСТНОСТИ ИНТЕРВЬЮ ПРИ ИНФОРМАЦИОННОМ ДЕФИЦИТЕ

*Суворова А.В., Лавренов А.В., Тулупьева Т.В., Тулупьев А.Л., Пащенко А.Е.* **Моделирование социально-значимого поведения респондентов: аналитическая и численная оценки интенсивности в окрестности интервью при информационном дефиците.**
**Аннотация.** Рассматривается подход к улучшению процедур построения оценок различных параметров поведения респондентов по сведениям о его последних эпизодах, предшествующих интервью. Предложены методы обработки неопределенности исходных данных, основанные на смешанном вероятностно-нечетком подходе. Получены аналитические, включая их асимптотические, приближения и численные оценки интенсивности поведения. Разработаны программные приложения, обеспечивающие возможность проведения численных экспериментов, реализующих предложенные процедуры обработки.
**Ключевые слова:** модели поведения, рискованное поведение, неопределенность, систематическая ошибка, последние эпизоды.

*Suvorova A.V., Lavrenov A.V., Tulupyeva T.V., Tulupyev A.L., Paschenko A.E.* **Modeling of socially significant respondents' behavior: analytical and numerical rate estimates based on the episodes near interview in case of information deficiency.**
**Abstract.** The paper offers a method for respondents' behavior modeling based on data about last episodes adjacent to interview and proposes improved techniques of modeling and processing of initial data uncertainty based on hybrid probabilistic and fuzzy approaches. This paper provides analytical (including asymptotic) analysis of these estimates and numerical behavior rate estimates according to the model. Software supplements enabling to fulfill numerical experiments realizing the proposed processing procedures were worked out.
**Keywords:** behavior models, risky behavior, uncertainty, bias, last episodes.

**1. Введение.** Задачи моделирования социально-значимого поведения респондентов, оценки его интенсивности, последующего расчета производных показателей (например, кумулятивного риска, ожидаемого или предотвращенного ущерба) возникают во многих отраслях социологических, психологических, эпидемиологических, экономико-демографических, социотехнических [1–3] исследований. Например, в эпидемиологии важна оценка риска передачи или приобретения неизлечимых инфекций (в частности, ВИЧ), а для этого необходимо знать интенсивность рискованного поведения респондентов.

Таким образом, в связи с задачами своевременного обнаружения изменений в поведении отдельных индивидов и групп науки социогу-

манитарного цикла испытывают потребность в математических моделях и алгоритмах, которые позволили бы получать оценки интенсивности социально-значимого поведения. При этом существующие методы прямого измерения интенсивности (круглосуточный мониторинг, дневниковый метод, длительное сопровождение когорты индивидов и пр.) часто не применимы из-за их дороговизны, а также из-за ряда проблем этического характера.

Отметим, что наиболее доступными исходными данными для анализа поведения выступают самоотчеты респондентов об их поведении, то есть ответы в анкете на блок вопросов или результаты интервью. На данный момент разработаны и применяются в опросах два подхода к оцениванию интенсивности поведения: прямые вопросы и Лайкерт-шкалы — каждый из которых имеет недостатки [4]. Одной из возможных альтернатив представляется опрос респондента о нескольких последних эпизодах его поведения. Однако ограниченное число и неточность, фактически, *нечеткость* естественно-языковых формулировок ответов (например, «на прошлой неделе»), не позволяют напрямую использовать известные методы из теории массового обслуживания для оценки интенсивности поведения, поэтому возникает потребность в новых — гибридных — математических моделях.

В результате все более актуальной становится междисциплинарная фундаментальная научная проблема — развитие методологии поиска, представления, агрегирования и обработки данных и знаний (полученных из самоотчетов респондентов) в условиях информационного дефицита для последующего формирования и расчета косвенных оценок интенсивности социально-значимого поведения. Эта проблема требует развития моделей и алгоритмов в рамках специфических математических и компьютерных дисциплин: теории принятия решений, искусственного интеллекта, мягких вычислений, теории вероятностей и математической статистики. Коллектив авторов доклада уже имеет некоторые результаты в рамках указанной научной проблемы [4–8]; *целью* же настоящей работы является формирование корректной модели для представления интервала между двумя эпизодами социально значимого поведения, «прерванного» моментом интервью.

Поведение рассматривается как случайный процесс некоторого класса. При этом встают вопросы о том, как оцениваются параметры этого процесса, как осуществляется обработка неполных исходных данных.

**2. Описание подхода.** Рассмотрим используемый подход к анализу данных в условиях дефицита информации [9]. Более подробно он описан в [4–8].

В результате интервью становятся известными сведения о нескольких (в рассматриваемом случае — трех) последних эпизодах поведения. Серия эпизодов рассматривается как пуассоновский случайный процесс с основным уравнением [10]

$$\mathrm{P}\left[ N\left([t_0, t_0 + T]\right) = k \right] = \frac{e^{-\lambda t}(\lambda T)^k}{k!},$$

где $t_0$ — начальный момент времени, $k$ — число последовательных событий, которые вспомнил респондент, а $T$ — тот период времени, за который эти эпизоды произошли, $\lambda$ — интенсивность.

Цель исследований — определение или оценка величины параметра $\lambda$, характеризующего интенсивность участия респондента в поведении определенного вида, а также его производные характеристики.

Применив метод максимального правдоподобия [11] к основному уравнению пуассоновского процесса при вышеуказанных данных, получим оценку интенсивности $\hat{\lambda} = \frac{k}{T}$.

Однако рассмотренный подход оставляет без внимания некоторые детали.

Отметим, что для пуассоновского процесса нам известна функция плотности распределения длин интервалов между соседними эпизодами поведения $p(\tau) = \lambda e^{-\lambda\tau}$, где $\lambda$ — интенсивность. В нашем случае величина $\tau$ отвечает длине интервала между эпизодами.

В рассматриваемой нами задаче обозначим через $\tau_0$ длину интервала между моментом интервью и последним эпизодом, $\tau_1$ — интервал между последним и предпоследним эпизодами, $\tau_2$ — между предпоследним и третьим с конца, $T$ — между моментом интервью и третьим с конца эпизодом.

При оценивании по методу максимального правдоподобия мы неявно предполагаем, что в день интервью также произошёл эпизод поведения (поскольку мы относимся к интервалу $\tau_0$ как к интервалу между эпизодами). Таким образом, полученная выше оценка является оценкой сверху для интенсивности. Обозначим в связи с этим

$$\lambda_{\max} = \frac{3}{T}. \qquad (1)$$

**3. Коррекция подхода к учёту последнего интервала.** Как было показано в предыдущем разделе, при использовании существующего подхода к анализу данных, полученных в результате интервью, делается необоснованное предположение о том, что в момент интервью также происходит эпизод поведения [12].

Введём дополнительные обозначения. Если $x$ — случайная величина с оговоренным распределением, то через $p(x)$ обозначим значение функции плотности $x$. Через $\pi(x)$ обозначим частный случай — экспоненциальное распределение $\lambda e^{-\lambda x}$. Тогда для случайного интервала $T$, подчиняющегося экспоненциальному распределению с интенсивностью $\lambda$, $p(T) = \pi(T)$. Пусть $\tau$ — случайная величина, отвечающая прерыванию интервала $T$ моментом интервью — имеет функцию плотности $p_T(\tau)$. Предположим (и это предположение оправдано, поскольку «жизненные планы» интервьюера и респондента независимы), что эти случайные величины являются независимыми. Тогда для случайной величины $\tau_0$, соответствующей интервалу между последним эпизодом поведения и моментом интервью, функция плотности имеет вид $p(\tau_0) = \int_0^\infty p_T(\tau_0)\, p(T) dT$.

Предполагая, что интервью может случиться в любой момент интервала $[0, T]$ с равной вероятностью, получим

$$p(\tau_0) = \int_{\tau_0}^{\infty} \frac{\lambda e^{-\lambda T}}{T} dT.$$

Обозначим через $\varpi(x) = \varpi_\lambda(x)$ выражение $\int_x^\infty \frac{\lambda e^{-\lambda u}}{u} du$. Таким образом, $p(\tau_0) = \varpi(\tau_0)$. По принципу максимального правдоподобия:

$$\tilde{\lambda} = \arg\max P(\lambda) = \arg\max\, p(\tau_0)\, p(\tau_1)\, p(\tau_2) =$$

$$= \arg\max \varpi(\tau_0)\pi(\tau_1)\pi(\tau_2) = \arg\max\left( \lambda^3 e^{-\lambda(\tau_1+\tau_2)} \int_{\tau_0}^{\infty} \frac{e^{-\lambda u}}{u} du \right).$$

Поскольку

$$\frac{e^{-x}}{x} = \int_x^\infty \frac{e^{-u}}{x} du > \int_x^\infty \frac{e^{-u}}{u} du, \qquad (2)$$

то $P(\lambda) \to 0$ при $\lambda \to \infty$; кроме того, $P(0) = 0$, эта функция неотрицательна и принимает положительные значения. Таким образом, для нахождения максимума $P(\lambda)$ достаточно исследовать ненулевые экстремумы этой функции.

$$\frac{dP}{d\lambda} = 0,$$

$$\lambda^2 e^{-\lambda(\tau_1+\tau_2)} \left( 3\int_{\tau_0}^{\infty} \frac{e^{-\lambda u}}{u} du - \lambda(\tau_1+\tau_2)\int_{\tau_0}^{\infty} \frac{e^{-\lambda u}}{u} du - e^{-\lambda\tau_0} \right) = 0,$$

$$3 - \lambda(\tau_1+\tau_2) - \frac{e^{-\lambda\tau_0}}{\left( \int_{\tau_0}^{\infty} \frac{e^{-\lambda u}}{u} du \right)} = 0. \qquad (3)$$

Далее будем ссылаться на уравнение (3) как на *уравнение интенсивности*. Следующий раздел будет посвящён исследованию этого уравнения.

**3. Исследование уравнения интенсивности.** Рассмотрим функцию

$$F(\lambda) = 3 - \lambda(\tau_1+\tau_2) - \frac{e^{-\lambda\tau_0}}{\int_{\tau_0}^{\infty} \frac{e^{-\lambda u}}{u} du}. \qquad (4)$$

Ясно, что $F(\lambda_{\max})$ (см. (1)) с использованием (2) можно оценить следующим образом:

$$F(\lambda_{\max}) < 3 - \lambda_{\max}(\tau_1+\tau_2) - \lambda_{\max}\tau_0 = 0.$$

С другой стороны, при $\lambda \to 0$ $F(\lambda) = 3 + o(1)$, в частности, при достаточно маленьком $\lambda$, $F(\lambda) > 0$. Таким образом, у уравнения $F(\lambda) = 0$ существуют решения.

Теперь исследуем $F(\lambda)$ на монотонность. Для этого рассмотрим уравнение

$$\frac{dF}{d\lambda} = -(\tau_1+\tau_2) + \frac{\tau_0 e^{-\lambda\tau_0}}{\int_{\tau_0}^{\infty} \frac{e^{-\lambda u}}{u} du} - \frac{e^{-2\lambda\tau_0}}{\lambda\left( \int_{\tau_0}^{\infty} \frac{e^{-\lambda u}}{u} du \right)^2} = 0.$$

Это уравнение можно переписать следующим образом:

$$\left(\frac{e^{-\lambda\tau_0}}{\int\limits_{\tau_0}^{\infty}\frac{e^{-\lambda u}}{u}du}\right)^2-\lambda\tau_0\left(\frac{e^{-\lambda\tau_0}}{\int\limits_{\tau_0}^{\infty}\frac{e^{-\lambda u}}{u}du}\right)+\lambda(\tau_1+\tau_2)=0.$$

И если это уравнение имеет решения, то можно написать, что

$$\frac{e^{-\lambda\tau_0}}{\int\limits_{\tau_0}^{\infty}\frac{e^{-\lambda u}}{u}du}=\frac{1}{2}\left(\lambda\tau_0\pm\sqrt{\lambda^2\tau_0^2-4\lambda\left(\tau_1+\tau_2\right)}\right).$$

Но правая часть равенства меньше, чем $\lambda\tau_0$, а левая - строго больше (см. (2)), что невозможно. Таким образом, производная функции $F(\lambda)$ не обращается в ноль, то есть функция $F$ монотонна. Из этого можно заключить, что уравнение $F(\lambda)=0$ имеет единственное решение.

Так как при близких к нулю значениях $\lambda$ значение $F(\lambda)>0$, а $F(\lambda_{\max})<0$, то из монотонности $F(\lambda)$ сразу следует, что корень уравнения $F(\lambda)=0$ меньше $\lambda_{\max}$, то есть, что искомое значение интенсивности действительно меньше максимального.

Теперь, пользуясь тем, что

$$\int\limits_{\tau_0}^{\infty}\frac{e^{-\lambda u}}{u}du=E_1(\lambda\tau_0)=-\gamma-\ln(\lambda\tau_0)+\sum_{k=1}^{\infty}\frac{(-1)^{k+1}(\lambda\tau_0)^k}{k\cdot k!}. \qquad (5)$$

(где $\gamma$ — это постоянная Эйлера, $\gamma\approx 0{,}5772156649$) и заменяя сумму ряда на частичную сумму из достаточно большого числа слагаемых, искомое значение интенсивности можно найти с помощью компьютерных приближений с любой точностью. В конце этого раздела мы вернёмся к описанию такой программы, а сейчас обратим внимание на два важных частных случая.

Во-первых, рассматривая различные примеры в первом параграфе, мы обращали особенное внимание на случай, когда длина интервала $\tau_0$ между последним эпизодом и моментом интервью значительно меньше, чем длины интервалов $\tau_1$ и $\tau_2$. Поэтому интересно знать асимптотику $\lambda$ при $(\tau_1+\tau_2)\to\infty$. Во-вторых, не меньший интерес представляет асимптотика $\lambda$, когда, наоборот, $\tau_0\to\infty$. Ниже разберём два этих случая.

Пусть $(\tau_1+\tau_2)\to\infty$. Отметим, что тогда $\lambda<\lambda_{\max}\to 0$. И пусть

$$3-\lambda(\tau_1+\tau_2)-\frac{e^{-\lambda\tau_0}}{\int\limits_{\tau_0}^{\infty}\frac{e^{-\lambda u}}{u}du}=0.$$

Обозначим $z=\lambda\tau_0;\ c=\dfrac{\tau_1+\tau_2}{\tau_0}$.

Таким образом, уравнение интенсивности (3) перепишется как

$$3-cz-\frac{e^{-z}}{E_1(z)}=0 \qquad (6)$$

(см. (5)), где $c\to\infty$. Теперь найдём первые несколько слагаемых асимптотики $z$ при бесконечно больших $c$. В первом приближении можно записать, что $3-cz+o(1)=0$ или $z=\dfrac{3}{c}+o(c^{-1})$.

Можно вычислить ещё несколько слагаемых асимптотики:

$$z=\frac{3}{c}-\frac{1}{c\ln c}-\frac{\gamma+\ln 3}{c\ln^2 c}+\left(\frac{1}{c\ln^2 c}\right)$$

или

$$\lambda=\frac{3}{\tau_1+\tau_2}-\frac{1}{(\tau_1+\tau_2)\ln\left(\dfrac{\tau_1+\tau_2}{\tau_0}\right)}-\frac{\gamma+\ln 3}{(\tau_1+\tau_2)\ln^2\left(\dfrac{\tau_1+\tau_2}{\tau_0}\right)}+$$

$$+o\left(\frac{1}{(\tau_1+\tau_2)\ln^2(\tau_1+\tau_2)}\right).$$

Обозначим

$$\lambda_{\tau_1+\tau_2}=\frac{3}{\tau_1+\tau_2}-\frac{1}{(\tau_1+\tau_2)\ln\left(\dfrac{\tau_1+\tau_2}{\tau_0}\right)}-\frac{\gamma+\ln 3}{(\tau_1+\tau_2)\ln^2\left(\dfrac{\tau_1+\tau_2}{\tau_0}\right)}. \qquad (7)$$

В конце параграфа мы отдельно обсудим, насколько хороши это и другие приближения, поэтому в данный момент оставим этот вопрос без внимания.

Теперь пусть $\tau_0\to\infty$. Снова рассмотрим уравнение (6), однако в нём теперь $c\to\infty$, и вычислим первые несколько слагаемых асимптотики $z$. Заметим, что, хотя $\lambda\to\infty$, $z$ не обязана быть бесконечно малой величиной, однако, поскольку $\lambda<\lambda_{\max}$, ясно, что величина $z$

ограничена. Пусть $z \to \xi < \infty$. Тогда, раскладывая функции $e^{-z}$ и $E_1(z)$ в ряд в точке $\xi$, в первом приближении получаем:

$$3 + o(1) - \frac{e^{-\xi} + o(1)}{E_1(\xi) + o(1)} = 0,$$

то есть $\xi$ должно быть решением уравнения $3 - \dfrac{e^{-\xi}}{E_1(\xi) + o(1)} = 0$. Поскольку это частный случай уравнения $F(\xi) = 0$ (см.(4)), существует единственное решение такого уравнения, и вычисления показывают, что $\xi \approx 2,2196304088$.

Таким образом, $z = \xi + o(1)$, то есть $\lambda = \dfrac{\xi}{\tau_0} + o\left(\dfrac{1}{\tau_0}\right)$.

Можно вычислить ещё несколько слагаемых асимптотики:

$$z = \xi + \frac{\xi^2}{3(\xi - 3)} c + \frac{\xi^3(\xi^2 - 7\xi + 9)}{18(\xi - 3)^3} c^2 + o(c^2)$$

или

$$\lambda = \xi\tau_0^{-1} + \frac{\xi^2(\tau_1 + \tau_2)}{3(\xi - 3)}\tau_0^{-2} + \frac{\xi^3(\xi^2 - 7\xi + 9)(\tau_1 + \tau_2)^2}{18(\xi - 3)^3}\tau_0^{-3} + o(\tau_0^{-3}). \qquad (8)$$

Наконец, вернёмся к численным приближениям решения уравнения интенсивности. Для этой цели была разработана программа, которая по значениям $\tau_1 + \tau_2$ и $\tau_0$ вычисляет значение интенсивности $\lambda$ (с точностью более чем $10^{-10}$), также для сравнения приводит величины $\lambda_{\max}$, $\lambda_{\tau_1+\tau_2}$, $\lambda_{\tau_0}$ (см. (1), (7), (8)) и, кроме того, приводит значение выражения $F(\lambda)$ (см.(4)), где $\lambda$ — приближенно вычисленное значение интенсивности.

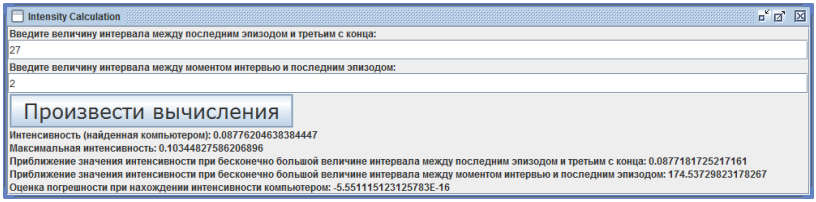


Рис. 1. Интерфейс программы, вычисляющей интенсивность

В таблице приведены значения $\lambda$ и $\lambda_{\max}$, $\lambda_{\tau_1+\tau_2}$, $\lambda_{\tau_0}$ на ряде примеров. Полужирным шрифтом выделены совпадающие знаки после запятой для первоначальной, уточненной и асимптотических оценок. К сожалению, вычисленные приближения $\lambda_{\tau_1+\tau_2}$ и $\lambda_{\tau_0}$, хотя и ведут себя несколько лучше, но оказываются мало применимы в реальных исследованиях, когда величины интервалов, зачастую, мало отличаются друг от друга. Таким образом, в приложениях, не требующих большой точности, можно использовать простую формулу вида (1). В общем случае исправленная оценка более точна.

Таблица. **Сравнение оценок интенсивности**

| $\tau_1+\tau_2$ | $\tau_0$ | $\lambda$ | $\lambda_{\max}$ | $\lambda_{\tau_1+\tau_2}$ | $\lambda_{\tau_0}$ |
|---|---|---|---|---|---|
| 100.0 | 20.0 | 0.0203610 | **0.02**5 | **0.0**17316997 | 2.158702461 |
| 100.0 | 1.0 | 0.0268266 | **0.02**97029702 | **0.02**7038325 | 20382.45253 |
| 80.0 | 1.0 | 0.0332707 | **0.03**70370370 | **0.033**556527 | 13011.89740 |
| 60.0 | 1.0 | 0.0438485 | **0.04**91803278 | **0.04**4263212 | 7288.596546 |
| 40.0 | 5.0 | 0.0553751 | **0.0**666666666 | **0.05**3288607 | 23.43286575 |
| 40.0 | 1.0 | 0.0644782 | **0.0**731707317 | **0.06**5144081 | 3212.549964 |
| 20.0 | 3.0 | 0.1075219 | **0.1**304347826 | **0.10**0362820 | 26.56801792 |
| 20.0 | 1.0 | 0.1230022 | **0.1**428571428 | **0.123**972889 | 783.7576592 |
| 10.0 | 2.0 | 0.2036110 | **0.2**5 | **0.**173169978 | 21.58702461 |
| 10.0 | 1.0 | 0.2286214 | **0.2**727272727 | **0.22**4962476 | 187.0818603 |
| 2.0 | 5.0 | 0.3228036 | **0.**4285714285 | 1.047673035 | **0.3**41459815 |
| 5.0 | 10.0 | 0.1512060 | **0.**2 | **0.1**90935744 | **0.1**68216963 |
| 2.0 | 15.0 | 0.1314253 | **0.1**764705882 | 1.541759718 | **0.131**709466 |
| 5.0 | 15.0 | 0.1126843 | **0.1**5 | **0**.504351271 | **0.11**6462038 |

**5. Обработка неопределенности, обусловленной нечеткостью ответов респондентов.** Как было отмечено ранее, ответы на вопросы об эпизодах поведения поступают на естественном языке, т.е. являются в значительной степени нечеткими и неполными. Отметим, что респонденты используют в своих высказываниях разные единицы измерения: часы, дни, недели, месяцы, полугода, года. Причем использованная единица измерения несет в себе информацию о точности измерения. Поясним это на примере двух, на первый взгляд, равнозначных высказываний: «семь дней назад» и «неделю назад». Когда респондент использует формулировку «семь дней назад», это свидетельствует о его уверенности в том, что событие произошло именно семь дней назад. В то время как «неделю назад» — это может быть и пять, и восемь дней назад.

Для учета указанной неточности каждый ответ рассматривается не как точка на временной оси, а как интервал, длина которого зависит от единицы измерения. Значение каждого ответа рассматривается, таким

образом, не как константа, а как случайная величина с заранее заданным распределением [4]. Введенная случайная величина за счет рандомизации [9] неопределенности ответа, обусловленной нечеткостью его формулировки, позволяет рассмотреть интенсивность как случайную величину и вычислить характеристики последней. Для каждого значения соответствующий интервал разбивается на $m$ частей. Рассматриваются все возможные сочетания точек из интервалов, вычисляются их веса и рассчитываются значения интенсивности по описанному выше методу. Затем строится обобщенная взвешенная оценка [9].

**6. Заключение**. В данной работе рассматривается один из способов модернизации подхода к описанию социально-значимого поведения респондента по данным о его последних эпизодах. Описанный метод позволяет избежать неявного предположения о том, что в момент интервью происходит следующий эпизод. Другой подход, учитывающий особенности интервала между последним эпизодом и моментом интервью, рассматривает введение распределения особого вида (принадлежащего к классу beta-prime — простых бета- — распределений) [13].

Однако результатов, полученных в рамках классических операций с распределениями вероятности, моментами случайных величин и функциями правдоподобия, оказывается недостаточно при переходе к неточным данным, доступным в результате проведения интервьюирования или опроса респондентов. То есть, для дальнейшей обработки неопределенности, связанной с нечеткостью исходных данных, требуется использовать метод сводных показателей [9]. Другим возможным решением является анализ рассмотренных случайных величин в рамках подхода, связанного с вероятностно-графическими моделями [14, 15].

## Литература


1. *Тулупьев А.Л., Азаров А.А., Тулупьева Т.В., Пащенко А.Е., Степашкин М.В.* Социально-психологические факторы, влияющие на степень уязвимости пользователей автоматизированных информационных систем с точки зрения социоинженерных атак // Труды СПИИРАН. 2010. Вып. 1 (12). С. 200–214.
2. *Ванюшичева О.Ю., Тулупьева Т.В., Пащенко А.Е., Тулупьев А.Л., Азаров А.А.* Количественные измерения поведенческих проявлений уязвимостей пользователя, ассоциированных с социоинженерными атаками. // Труды СПИИРАН. 2011. Вып. 19. С. 34–47.
3. *Тулупьева Т.В., Тулупьев А.Л., Азаров А.А., Пащенко А.Е.* Психологическая защита как фактор уязвимости пользователя в контексте социоинженерных атак // Труды СПИИРАН. 2011. Вып. 18. С. 74–92.

4. *Суворова А.В., Тулупьев А.Л., Пащенко А.Е., Тулупьева Т.В., Красносельских Т.В.* Анализ гранулярных данных и знаний в задачах исследования социально значимых видов поведения // Компьютерные инструменты в образовании. №4. 2010. С. 30–38.
5. *Тулупьева Т.В., Пащенко А.Е., Тулупьев А.Л., Красносельских Т.В., Казакова О.С.* Модели ВИЧ-рискованного поведения в контексте психологической защиты и других адаптивных стилей. СПб.: Наука, 2008. 140 с.
6. *Пащенко А. Е., Тулупьев А. Л., Николенко С. И.* Моделирование заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения. // Известия высших учебных заведений: Приборостроение. 2006. №8. 33–34 с.
7. *Тулупьева Т.В., Тулупьев А.Л., Пащенко А.Е.* Оценка интенсивности поведения респондента в условиях информационного дефицита // Труды СПИИРАН. Вып. 7. СПб.: Наука, 2008. С. 239–254.
8. *Тулупьева Т.В., Пащенко А.Е., Тулупьев А.Л., Голянич В.М.* Модели ВИЧ-рискованного поведения в контексте психологической защиты и адаптации // Вестник СПбГУ. 2010. Серия 12. Вып. 1. С. 95–104.
9. *Хованов Н.В.* Анализ и синтез показателей при информационном дефиците. СПб.: Изд-во СПбГУ, 1996. 196 с.
10. *Розанов Ю.А.* Случайные процессы (краткий курс). М.: Главная редакция физико-математической литературы издательства «Наука», 1971. 288 с.
11. *Крамер Г.* Математические методы статистики. М.: Мир, 1975. 648 с.
12. *Лавренов А.В., Суворова А.В., Пащенко А.Е., Тулупьев А.Л.* Особенности обработки данных и знаний об эпизодах социально-значимого поведения в окрестности интервью // Труды СПИИРАН. 2010. Вып. 15. С. 246–262;
13. *Зельтерман Д., Тулупьев А.Л., Суворова А.В., Пащенко А.Е., Мусина В.Ф., Тулупьева Т.В., Красносельских Т.В., Гро Л., Хаймер Р.* Обработка систематической ошибки, связанной с длиной временных интервалов между интервью и последним эпизодом в гамма-пуассоновской модели поведения // Труды СПИИРАН. 2011. Вып. 16. С. 160–185.
14. *Фильченков А.А., Тулупьев А.Л.* Совпадение множеств минимальных и нередуцируемых графов смежности над первичной структурой алгебраической байесовской сети // Вестник Санкт-Петербургского государственного университета. Серия 1. Математика. Механика. Астрономия. 2012. Вып. 2. С. 65–74.
15. *Тулупьев А.Л., Фильченков А.А., Вальтман Н.А.* Алгебраические байесовские сети: задачи автоматического обучения // Информационно-измерительные и управляющие системы. 2011. № 11, т. 9. С. 57-61.

**Суворова Алена Владимировна** — младший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики СПИИРАН, аспирант математико-механического факультета Санкт-Петербургского государственного университета (СПбГУ). Область научных интересов: математическая статистика, теория вероятности, применение методов математического моделирования в эпидемиологии. Число научных публикаций — 21. SuvorovaAV@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450. Научный руководитель — А.Л. Тулупьев.

**Suvorova Alena Vladimirovna** — junior researcher, Laboratory of Theoretical and Interdisciplinary Computer Science, SPIIRAS, PhD student, Faculty of Mathematics and Mechanics of

St. Petersburg State University (SPbSU). Research interests: mathematical statistics, probability theory, application of mathematical modeling in epidemiology. The number of publications — 21. SuvorovaAV@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450. Scientific advisor — A.L. Tulupiev.

**Лавренов Андрей Валентинович** — студент математико-механического факультета Санкт-Петербургского государственного университета (СПбГУ). Область научных интересов: математическая статистика, теория вероятности. Число научных публикаций — 2. vedrfiolnir@gmail.com; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450. Научный руководитель — А.Л. Тулупьев.

**Lavrenov Andrey Valentinovich** —student, Faculty of Mathematics and Mechanics of St. Petersburg State University (SPbSU). Research interests: mathematical statistics, probability theory. The number of publications — 2. vedrfiolnir@gmail.com; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450. Scientific advisor — A.L. Tulupiev.

**Тулупьева Татьяна Валентиновна** — канд. психол. наук, доцент; старший научный сотрудник лаборатории теоретических и междисциплинарных проблем информатики Учреждения Российской академии наук С.-Петербургский институт информатики и автоматизации РАН (СПИИРАН), доцент кафедры информатики математико-механического факультета С.-Петербургского государственного университета (СПбГУ), доцент кафедры психологии управления и педагогики Северо-Западной академии государственной службы (СЗАГС). Область научных интересов: применение методов математики и информатики в гуманитарных исследованиях, информатизация организации и проведения психологических исследований, применение методов биостатистики в эпидемиологии, психология личности, психология управления. Число научных публикаций — 70. TVT@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450.

**Tulupyeva Tatiana Valentinovna** — PhD in Psychology, associate professor; senior researcher, Laboratory of Theoretical and Interdisciplinary Computer Science, SPIIRAS, associate professor, Computer Science Department, Faculty of Mathematics and Mechanics, St. Petersburg State University (SPbSU), associate professor, Management Psychology and Pedagogic Department, North-West Academy of Public Administration (NWAPA). Research interests: application of mathematics and computer science in humanities, informatization of psychological studies, application of biostatistics in epidemiology, psychology of personality, management psychology. Number of publications — 70. TVT@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450.

**Тулупьев Александр Львович** — д-р физ.-мат. наук, доцент; заведующий лабораторией теоретических и междисциплинарных проблем информатики СПИИРАН, профессор кафедры информатики математико-механического факультета С.-Петербургского государственного университета (СПбГУ). Область научных интересов: представление и обработка данных и знаний с неопределенностью, применение методов математики и информатики в социокультурных исследованиях, применение методов биостатистики и математического моделирования в эпидемиологии, технология разработки программных

комплексов с СУБД. Число научных публикаций — 210. ALT@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450.

**Tulupyev Alexander Lvovich** — PhD in Appl. Math. and CS, Dr. Sci. in CS, associate professor; head of laboratory, Theoretical and Interdisciplinary Computer Science Laboratory, SPIIRAS, professor, Computer Science Department, Faculty of Mathematics and Mechanics, St. Petersburg State University (SPbSU). Research interests: uncertain knowledge and data representation and processing, application of mathematics and computer science in sociocultural studies, applications of biostatistics and mathematical modeling in modern epidemiology, software technologies and development of information systems with databases. The number of publications — 210. ALT@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)3284450.

**Пащенко Антон Евгеньевич** — н. с. лабораторией теоретических и междисциплинарных проблем информатики СПИИРАН. Область научных интересов: математическая статистика, статистическое моделирование, применение методов биостатистики и математического моделирования в эпидемиологии. Число научных публикаций — 35. AEP@iias.spb.su, www.tulupyev.spb.ru; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-3337, факс +7(812)328-4450.

**Paschenko Anton Evgen'evich** — researcher, Theoretical and Interdisciplinary Computer Science Laboratory, SPIIRAS. Research interests: mathematical statistics, statistical modeling, application of biostatistics and mathematical modeling in epidemiology. The number of publications — 35. AEP@iias.spb.su, www.tulupyev.spb.ru; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-3337, fax +7(812)328-4450.

Рекомендовано ТиМПИ СПИИРАН, зав. лаб. д-р физ.-мат. наук, доцент А.Л. Тулупьев.
Статья поступила в редакцию 23.03.2012.

## РЕФЕРАТ

*Суворова А.В., Лавренов А.В., Тулупьева Т.В., Тулупьев А.Л., Пащенко А.Е.* **Моделирование социально-значимого поведения респондентов: аналитическая и численная оценки интенсивности в окрестности интервью при информационном дефиците.**

Задачи оценивания интенсивности социально-значимого поведения респондентов по их самоотчетам об эпизодах поведения возникают во многих отраслях социологических, психологических, маркетинговых исследований.

Заметим, что ответы респондента на вопросы о последних эпизодах характеризуются стабильностью воспроизведения. Однако ограниченное число и неточность, недоопределенность, нечеткость естественно-языковых формулировок ответов (то есть наблюдаемый сверхкороткий временной ряд) не позволяют напрямую использовать известные методы из теории массового обслуживания для оценки интенсивности поведения, поэтому возникает необходимость в предложении новых математических моделей.

Поведение рассматривается как случайный процесс некоторого класса. При этом встают вопросы о том, какой процесс лучше описывает поведение, как меняются параметры этого процесса, как осуществляется обработка неполных исходных данных. Цель данной статьи — описать проблемы, возникающие при анализе данных о последних эпизодах социально-значимого поведения, и предложить некоторые пути их решения.

Используемые в настоящее время методы нахождения интенсивности, несмотря на логичные и правомерные мотивирующие соображения, имеют ряд недостатков из-за некоторых оставленных без внимания деталей. В частности, описанный в данной работе метод позволяет избежать неявного предположения о том, что в момент интервью происходит следующий эпизод. Получены аналитические, включая их асимптотические приближения, и численные оценки интенсивности поведения.

Предложены методы обработки неопределенности исходных данных, основанные на смешанном вероятностно-нечетком подходе. Разработаны программные приложения, обеспечивающие возможность проведения численных экспериментов, реализующих предложенные модели обработки.

## SUMMARY

*Suvorova A.V., Lavrenov A.V., Tulupyeva T.V., Tulupyev A.L., Paschenko A.E.*
**Modeling of socially significant respondents' behavior: analytical and numerical rate estimates based on the episodes near interview in case of information deficiency.**

We are faced with the problem of socially significant behavior rate estimate on the base of respondents' self-reports about their behavior episodes in many fields of sociological, psychological and marketing research.

Note that respondents' answers about last behavior episodes are stable. Limited number of natural language responses' forms and their fuzziness, uncertainty, imprecision (or super-short time series) do not allow to use directly the known methods of queuing theory for behavior rate estimation, that's why we need to propose new mathematical behavior models.

The behavior is described by a random process. And we have to find what random process type is the best for describing behavior, how the parameters of this process change, how incomplete data should be handled. The aim of this paper is to describe problems arising in the process of analysis of data about the last episodes of socially significant behavior and to propose several ways of their solution.

The used nowadays methods of rate finding despite their logical and valid motivating arguments have several drawbacks because of some unnoticed details. In particular, our method excludes the assumption that one more behavior episode takes place at the time of interview. This paper provides analytical (including asymptotic analysis of these estimates) and numerical behavior rate estimates according to the model.

The paper proposes improved techniques of modeling and processing of initial data uncertainty based on hybrid probabilistic and fuzzy approaches. The developed software applications allow to fulfill numerical experiments realizing the proposed processing models.