

УДК 519.254

И.В. Горошко, Ш.Х. Гонов  
**РАЗРАБОТКА АЛГОРИТМА ОЦЕНКИ РЕЗУЛЬТАТОВ  
ДЕЯТЕЛЬНОСТИ ОРГАНОВ ВНУТРЕННИХ ДЕЛ С  
ИСПОЛЬЗОВАНИЕМ МОДЕЛЕЙ БИНАРНОГО ВЫБОРА**  
*Академия управления МВД России*

*В статье предлагается технологический алгоритм анализа на основе моделей дискретного выбора, позволяющий оценивать оперативно-служебную деятельность органов внутренних дел в сфере обеспечения правопорядка в общественных местах. Предметом исследования являются методы и модели анализа результатов деятельности территориальных органов МВД России. В ходе исследования применялись методы многомерного статистического анализа данных, а также математический аппарат моделей дискретного выбора. Рассматривается механизм идентификации и оценки параметров logit- и probit-моделей на основе метода максимального правдоподобия. Предложены методы оценки значимости параметров регрессионных уравнений, а также методы оценки качества модели. Проведенный анализ единой совокупности статистических данных при помощи logit- и probit-моделей позволил сделать вывод о наибольшем влиянии оценочных показателей, имеющих более высокие коэффициенты значимости в системе оценки эффективности. Моделирование показало возможности применения данных моделей в поиске количественных закономерностей развития изучаемого процесса. Впервые с использованием математических моделей дискретного выбора разработан подход к анализу и оценке преступности в общественных местах, на их основе предложены новые решения по оценке результатов оперативно-служебной деятельности органов внутренних дел, что позволяет повысить качество информационно-аналитической работы.*

**Ключевые слова:** алгоритм, модели бинарного выбора, оценка, метод максимального правдоподобия, логит-модель.

**Введение.** Выработка и принятие эффективных мер по противодействию преступности является важнейшим приоритетом внутренней государственной политики. Вместе с тем, современные условия функционирования органов внутренних дел (далее ОВД) характеризуются нарастанием информационных потоков, содержание которых не всегда отвечает требованиям актуальности, полноты и достоверности, что затрудняет принятие эффективных решений и требует использования современных математических методов и моделей, основанных на получении, обработке и представлении сведений, используемых в оперативно-служебной деятельности [7].

Широкий спектр применения регрессионного анализа в социально-экономических исследованиях позволяет изучать влияние организационных, управленческих, природно-климатических и территориальных факторов на количественные изменения показателей

преступности, а также оценивать оперативно-служебную деятельность органов внутренних дел (далее ОВД).

**Постановка задачи.** В информационно-аналитической деятельности ОВД бинарные модели данных могут применяться как для анализа криминогенной обстановки, так и для оценки результатов деятельности ОВД [5]. Оценка результатов деятельности территориальных органов МВД России может быть представлена в виде переменной с тремя качественными индикаторами (*{«неудовлетворительно», «удовлетворительно», «отлично»}*) [10]. Для бинарных данных необходимо использование трех переменных, соответствующих каждой степени оценки:  $y^i=1$  неудовлетворительная,  $y^i=0$  остальные;  $y^g=1$  удовлетворительная,  $y^g=0$  остальные;  $y^t=1$  отличная,  $y^t=0$  остальные. Выбор альтернативы представлен в виде дерева, в узлах которой происходит выбор решения (Рисунок 1).

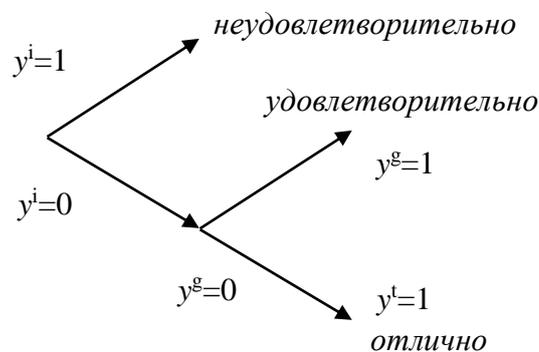


Рис. 1. Дерево решений.

В аналитической деятельности ОВД одной из важнейших стадий является оценка степени соответствия полученных результатов поставленным целям. Очевидно, что основным результатом деятельности подразделений полиции в сфере правопорядка в общественных местах является сохранение стабильной оперативной обстановки [9]. Под данным термином мы понимаем устойчивую динамику показателей преступности на определенной территории, в конкретных условиях внешней среды. По нашему мнению, использование одного единственного показателя не отражает реальную степень общественной опасности данного вида преступности. Решение нам видится в применении комплексного показателя, характеризующего различные стороны оперативной обстановки.

Для решения данной задачи нами предлагается следующий алгоритм оценки результатов деятельности территориальных органов МВД России.

**Алгоритм оценки.** Разработанный нами алгоритм состоит из 7 этапов:

1. Определение значений индикаторной переменной за счет применения механизмов комплексного оценивания.
2. Построение моделей бинарного выбора (*logit*, *probit*).
3. Оценка неизвестных параметров модели с использованием метода максимального правдоподобия. На основе статистики Вальда определяется уровень значимости коэффициентов модели.
4. Расчет и сравнение показателей качества модели: статистика отношения правдоподобия (*LR*), аналоги коэффициента детерминации ( $R^2_{pseudo}$ ,  $R^2_{McFadden}$ ,  $R^2_{C\&S}$ ,  $R^2_N$ ) и статистика Хосмера-Лемшоу.
5. Построение ROC-кривой для анализа доли корректных прогнозов.
6. Оценка связи между переменными путем расчета коэффициентов ( $\gamma$ -коэффициента Гудмана и Крускала,  $\tau$ - $a$ -коэффициента Кэндэла,  $D$ -коэффициента Сомерса);

Выбор наилучшей модели на основе информационных критериев: Акаике (AIC) и Бейесовского информационного критерия Шварца (BIC).

Рассмотрим механизм оценки с применением моделей бинарного выбора. Обозначим удовлетворительный уровень преступности  $y$ , и будем считать, что  $y = 1$ , если результаты раскрытия преступлений оставались стабильными. Набор множества факторов, оказывающих влияние на результаты можно представить вектором  $x = (x_1, \dots, x_k)'$ . Следовательно, модель бинарного выбора выглядит как:

$$P(y_t = 1) = F(x'_t \cdot \beta), \quad (1)$$

где  $F(\cdot)$  – функция распределения, область значений которой лежит в диапазоне  $\{0, 1\}$ . При использовании в качестве функции распределения  $F(\cdot)$  стандартного нормального распределения модель называется *probit*, а стандартного логистического распределения *logit* [4].

Функции стандартного логистического и стандартного нормального распределения представлены на рисунке (Рисунок 2):

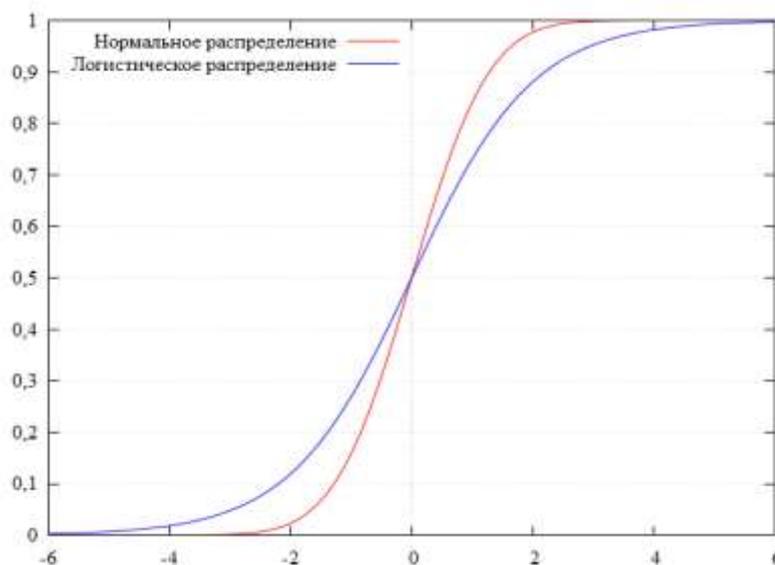


Рис. 2. Нормальное и логистическое распределение.

Для оценки коэффициентов  $\beta$  используем метод максимального правдоподобия (далее МПП) [11]. Для оценки параметров *logit*-модели рассчитаем функцию правдоподобия *PP*. Вероятность  $P(Y_i=y_i/X_{1i}, X_{2i}, \dots, X_{ki})$  – вероятность того, что предсказанная зависимая переменная  $y_i$  равна значению  $Y_i$  с учетом значений независимых переменных  $X_{1i}, X_{2i}, \dots, X_{ki}$ . В сокращенном виде данное выражение запишем как  $P(Y=y/X)$ . Вероятность  $P(Y=y/X)$  рассчитаем по формуле:

$$P(Y=y/X) = P(X)^y \cdot [1-P(X)]^{(1-y)} \quad (2)$$

Прологарифмировав обе части уравнения (2) получаем выражение следующего вида:

$$\ln[P(Y=y/X)] = y \cdot \ln[P(X)] + (1-y) \cdot \ln[[1-P(X)]] \quad (3)$$

Функция правдоподобия (*PP*) представляет собой сумму значений  $\ln[P(Y=y/X)]$  и рассчитывается по следующей формуле:

$$PP = \sum Y_i \cdot P(X_i) + (1-Y_i)(1-P(X_i)) \quad (4)$$

где  $P(x) = P(x) = \frac{e^L}{1 + e^L}$  .

Для получения оценок неизвестных параметров модели МПП следует максимизировать функцию правдоподобия. Полученное, в процессе реализации метода, нелинейное уравнение (4) может быть решено при помощи метода Ньютона, для которого требуется задание

начального значения коэффициентов  $\beta$  (0,01), максимального количества итераций и точности сходимости.

Оценку значимости коэффициентов построенной модели произведем на основе  $P$ -значения. Для этого рассчитаем стандартную ошибку и статистику Вальда. Стандартные ошибки коэффициентов равны квадратным корням диагонали ковариационной матрицы, на основе которых и рассчитывается статистика Вальда. Расчет осуществляется по формуле:

$$Wald = \left( \frac{\hat{\beta}}{SE} \right)^2 \quad (5)$$

где  $\beta$  – оцениваемые коэффициенты модели,  $SE$  – стандартная ошибка. Полученные значения в статистике Вальда выше 0,05 указывает статистическую незначимость коэффициентов.

На следующем этапе осуществляется тестирование значимости модели в целом. Для этого рассчитаем статистику отношения правдоподобия по формуле:

$$LR=2(l_1 - l_0), \quad (6)$$

где  $l_1, l_0$  – значения логарифмической функции правдоподобия. Если полученное значение больше, чем критическое значение распределения  $\chi^2(m)$ , где,  $m$  – число степеней свободы (независимых переменных), то модель признается значимой. Рассчитанные значения статистики правдоподобия также используются для расчета коэффициента детерминации. В рамках проведенного исследования нами применялось несколько аналогов классического коэффициента детерминации  $R^2$ :

- псевдо-коэффициент детерминации, рассчитываемый как:  $R^2_{pseudo} = LR/LR+n$ , где  $n$  – число наблюдений;
- индекс отношения правдоподобия МакФаддена, рассчитываемый по формуле:  $R^2_{McFadden} = 1 - l_1 / l_0$ ;
- индекс Кокса-Снелла, рассчитываемый по формуле:  $R^2_{C\&S} = 1 - (l_0 / l_1)^{2/n}$ . Ввиду того, что показатели  $l_1$  и  $l_0$  это значения логарифмической функции, обратной функцией экспоненцирования преобразуем их. Измененная формула выглядит как:  $R^2_{C\&S} = 1 - (exp(l_0) / exp(l_1))^{2/n}$ ;
- индекс Найджелкерка, рассчитываемый как:  $R^2_N = R^2_{C\&S} / 1 - l_0^{2/n}$ .

Все рассчитанные показатели используются для измерения силы связи между переменными, но в отличие от классического коэффициента детерминации их интерпретация является не всегда однозначной. Эти показатели принимают значение от 0 до 1, но обычно намного меньше коэффициента  $R^2$ , применяемого в обычном регрессионном анализе при

оценке тесноты связи по шкале Чеддока. Так, рассчитанное значение псевдокоэффициентов меньше 0,2 показывает слабую силу связи; от 0,2 до 0,4 демонстрирует среднюю силу; более 0,4 показывает сильную связь.

Показатели, применяемые для измерения силы связи между переменными мы будем рассматривать в комплексе с долей корректных прогнозов. Для данного анализа необходимо рассчитать статистику Хосмера-Лемшоу [2], для чего выборка разбивается на несколько подвыборок, по каждой из которых определяется доля данных со значением зависимой переменной равной 1:

$$p_j = \bar{y}_j = \sum_{i=1}^{n_j} y_{ij} / n_j \quad (7)$$

и предсказанная средняя вероятность по подгруппе:

$$\hat{p}_j = \sum_{i=1}^{n_j} \hat{p}_{ij} / n_j, \quad (8)$$

на основе которых и осуществляется расчет статистики *HL*:

$$HL = \sum_{i=1}^J \frac{n_j (p_j - \hat{p}_j)^2}{\hat{p}_j (1 - \hat{p}_j)}. \quad (9)$$

В случае если полученное значение *HL* больше, чем  $\chi^2(m - 2)$  то модель признается значимой. Для анализа доли корректных прогнозов целесообразно построение *ROC*-кривой [1]. В графическом виде *ROC*-кривая представлена на Рисунке 3.

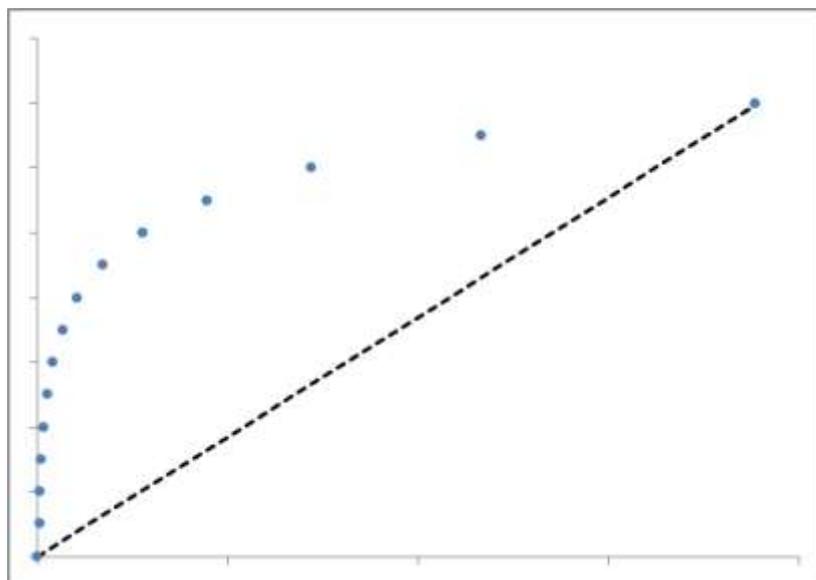


Рисунок 3 - *ROC*-кривая.

*ROC*-анализ позволяет оценить качество классификации и отображает соотношение между долями корректно предсказанных признаков (*TPR*):

$$FPR(\alpha, X^k) = \frac{\sum_{i=1}^k [\alpha(x_i) = +1][y_i = -1]}{\sum_{i=1}^k [y_i = -1]} \quad (10)$$

с признаками, ошибочно классифицированными, как несущими признак (*FPR*):

$$TPR(\alpha, X^k) = \frac{\sum_{i=1}^k [\alpha(x_i) = +1][y_i = +1]}{\sum_{i=1}^k [y_i = +1]} \quad (11)$$

Показатель *AUC* отображает количественную интерпретацию *ROC*-анализа и рассчитывается по формуле:

$$AUC = \int f(x)dx = \sum_i \left[ \frac{X_{i+1} + X_i}{2} \right] \cdot (Y_{i+1} - Y_i) \quad (12)$$

Чем выше данный показатель, тем качественнее построена модель.

Методы оценки взаимосвязи между переменными используются для исследования взаимосвязи между наблюдаемыми и прогнозируемыми значениями [3]. Для измерения силы связи между признаками произведем расчет некоторых показателей, по следующим формулам:

-  $\gamma$ -коэффициент Гудмана и Крускала:

$$G = \frac{N_s - N_d}{N_s + N_d}, \quad (13)$$

где  $N_t$  – число пар, связанных по зависимой переменной,  $N_s$  – число пар число пар случаев, ранжированных в том же порядке по обоим переменным (конкордантных пар), и  $N_d$  – число пар случаев, ранжированных в обратном порядке по обоим переменным. Эту статистику можно рассматривать как оценку максимального правдоподобия для теоретической величины  $\gamma$ , где  $\gamma = \frac{P_s - P_d}{P_s + P_d}$ .  $P_s$  и  $P_d$  – соответствующие вероятности того, что случайно выбранная пара наблюдений будет размещаться на таких же порядках при ранжировании по обоим переменным;

-  $\tau$ - $a$ -коэффициент Кэндэла:

$$\tau_a = \frac{N_s - N_d}{n(n-1)/2}, \quad (14)$$

-  $D$ -коэффициент Сомерса:

$$D_s = \frac{N_s - N_d}{N_s + N_d + N_t}, \quad (15)$$

где  $N_t$  – число пар, связанных по зависимой переменной. Высокие значения коэффициентов показывают, что модель обладает лучшей прогностической способностью.

На последнем этапе осуществляется расчет информационных критериев Акаике ( $AIC$ ) и Шварца ( $BIC$ ) по формулам:

$$AIC = n \ln \left( \frac{SSE}{n} \right) + 2k \quad (16)$$

$$BIC = n \ln \left( \frac{SSE}{n} \right) + k \ln(n) \quad (17)$$

где  $n$  – число наблюдений,  $k$  – число параметров модели,  $SSE$  (*Sum of squared errors*) – сумма квадратов отклонений.

Смысл данных критериев состоит в том, что модели, имеющие меньшее число параметров, являются более предпочтительными. При сравнении моделей лучшей признается та, для которой получено меньшее значение информационного критерия. Для применения в моделях бинарного выбора информационные критерии необходимо модернизировать.

$$AIC_{aug} = -2LL + 2k \quad (18)$$

$$BIC_{aug} = -2LL + k \ln(n) \quad (19)$$

где  $LL$  – логарифмическое правдоподобие.

**Основные результаты.** Построенная на основе разработанного алгоритма *logit*-модель выглядит следующим образом:

$$L = -12,1 + 0,311 \cdot TG\_RS + 0,101 \cdot LC\_PROP + 0,252 \cdot NL\_LICA, \quad R^2_{McFadden} = 0,55.$$

(2,77)      (0,114)      (0,0275)      (0,0900)

где  $TG\_RAS$  - количество раскрытых тяжких и особо тяжких преступлений общеуголовной направленности в расчете на 100 ед. штатной численности;

$LC\_PROP$  - число выявленных лиц по преступлениям против собственности, в расчете на 100 ед. штатной численности полиции;

$NL\_LICA$  - число несовершеннолетних, совершивших преступления в расчете на 1 тыс. несовершеннолетних в возрасте от 14 до 17 лет.

Проведенный анализ по единой совокупности статистических данных позволил сделать вывод о наибольшем влиянии тяжких и особо тяжких преступлений на уровень преступности в общественных местах. Так, результаты моделирования показали, что увеличение количества раскрытых данной категории приводит к улучшению оценочных показателей. Кроме того, была выявлена существенная значимость фактора выявленных лиц, совершивших преступления против собственности для

динамики преступлений, посягающих на общественный порядок и общественную безопасность. Заметим также, что показатель число несовершеннолетних, совершивших преступления, также является статистически значимым [6].

Следует также отметить полученный нами интересный результат, что такие, казалось бы, существенные показатели как, преступления, совершенные на бытовой почве и преступления, связанные с незаконным оборотом огнестрельного оружия, не оказывает статистически заметного влияния на оценку эффективности ОВД.

**Заключение.** Важным итогом проведенного исследования явилось обоснование возможности применения моделей бинарного выбора для анализа результатов оперативно-служебной деятельности и поиске количественных закономерностей развития изучаемого процесса. Следует отметить, что построенные logit- и probit-модели дают схожие результаты. Однако, выбор модели логистической регрессии является более предпочтительной по информационным критериям AIC и BIC.

Очевидно, что информационно-аналитическое обеспечение в сфере обеспечения правопорядка в общественных местах все еще имеет ряд нерешенных проблем. Решение этой задачи тесно связано с поиском перспективных подходов к оценке результатов деятельности ОВД, которым приходится работать в очень сложных условиях внешней среды. Без получения такой оценки руководителю трудно принимать эффективные управленческие решения, вырабатывать последовательность мероприятий, позволяющие подразделениям ОВД формировать оптимальные траектории своего развития.

#### ЛИТЕРАТУРА

1. Hanley J.A., McNeil B.J. The meaning and use the area under a receiver operating characteristic (ROC) Curve//Radiology 143(1). April, 1982. 29-36 pp.
2. Hosmer, D.W., Lemeshow S. Applied logistic regression. 2-nd ed. New York: John Wiley&Sons, 2000. – 375 p.
3. Roger Newson. Parameters behind «nonparametric» statistics: Kendall's tau, Somers' D and median differences. Stata Journal. 2001. 1 (1), 1–20 pp.
4. William H. Greene. Econometric Analysis. 7th ed. – Boston: Prentice Hall, 2012. – 1231 p.
5. Гонов Ш.Х. Моделирование влияния социально-экономических факторов на уровень преступности в общественных местах//Экономика и менеджмент систем управления. 2016. № 4(22). С. 74-80.

6. Гонов Ш.Х. Применение моделей бинарного выбора для оценки эффективности деятельности ОВД в сфере обеспечения правопорядка в общественных местах//Экономика и менеджмент систем управления. 2017. № 4(26). С. 70-78.
7. Горошко И.В., Лебедев А.В., Петрова В.Ю. Об одном из подходов к моделированию динамики преступности // Труды Академии управления МВД России. 2011. № 2. С. 54-58.
8. Горошко И.В. Математическое моделирование в управлении органами внутренних дел. – М.: Академия управления МВД России, 2000. – 145 с.
9. Демидов Ю.Н. О текущих задачах и проблемных аспектах в сфере организации деятельности подразделений полиции по охране общественного порядка территориальных органов МВД России/Ю. Н. Демидов // Научный портал МВД России, 2013, № 4. - С.104-110.
10. Клушин О.З. Оперативная обстановка: понятие, анализ, прогноз. – М.: Академия управления, 2010. – 256 с.
11. Носко В.П. Эконометрика для начинающих. - М.: ИЭПП, 2005. – 379 с.

I. V. Goroshko, S. H. Gonov

## **DEVELOPMENT OF ALGORITHM FOR ASSESSMENT OF RESULTS OF ACTIVITY OF POLICE DEPARTMENTS WITH THE USE OF BINARY CHOICE MODELS**

*Academy of Management of the Ministry of Internal Affairs of Russia*

*This paper proposes a technological algorithm for evaluating the results on the basis of discrete choice models, which allows to evaluate the activities of the police departments. The mechanism of identification and evaluation of the parameters of logit and probit models, as well as methods for assessing the quality and significance are considered. The possibility of application of these models for the analysis of results of activity of the police departments is proved. In the article the technological algorithm of the analysis on the basis of models of discrete choice allowing to estimate operational activity of the police departments in the sphere of ensuring a law and order in public places is offered. The subject of the research is the methods and models of analysis of the results of the activities of the police departments of the Ministry of internal Affairs of Russia. The methods of multidimensional statistical analysis of data, as well as the mathematical apparatus of discrete choice models were used in the study. The mechanism of identification and estimation of parameters of logit and probit models on the basis of the maximum likelihood estimation is considered. Proposed methods for evaluating the significance of the parameters of the regression equations, and methods of evaluating model quality. The analysis of statistical dataset using logit - and probit- models has led to the conclusion about the greatest impact of performance indicators with higher coefficients in the system of performance evaluation. Modeling showed the possibility of using these models in the search for quantitative laws of the studied process. For the first time,*

*using mathematical models of discrete choice, an approach to the analysis and evaluation of crime in public places was developed, on the basis of which new solutions were proposed to evaluate the results of operational and service activities of the police departments, which allows improving the quality of information and analytical work.*

**Keywords:** algorithm, model of binary choice, evaluation, maximum likelihood estimation, logit-model.

## REFERENCES

1. Hanley J.A., McNeil B.J. The meaning and use the area under a receiver operating characteristic (ROC) Curve//Radiology 143(1). April, 1982. 29-36 pp.
2. Hosmer, D.W., Lemeshow S. Applied logistic regression. 2-nd ed. New York: John Wiley&Sons, 2000. – 375 p.
3. Roger Newson. Parameters behind «nonparametric» statistics: Kendall's tau, Somers' D and median differences. Stata Journal. 2001. 1 (1), 1–20 pp.
4. William H. Greene. Econometric Analysis. 7th ed. – Boston: Prentice Hall, 2012. – 1231 p.
5. Gonov Sh.H. Modelirovanie vliyaniya social'no-ehkonomicheskikh faktorov na uroven' prestupnosti v obshchestvennykh mestah. Ekonomika i menedzhment sistem upravleniya. 2016. No. 4(22). pp. 74-80.
6. Gonov Sh.H. Primenenie modelej binarnogo vybora dlya ocenki ehffektivnosti deyatel'nosti OVD v sfere obespecheniya pravoporyadka v obshchestvennykh mestah. Ekonomika i menedzhment sistem upravleniya. 2017. No. 4(26). pp. 70-78.
7. Goroshko I.V., Lebedev A.V., Petrova V.Y. Ob odnom iz podhodov k modelirovaniyu dinamiki prestupnosti. Trudy Akademii upravleniya MVD Rossii. 2011. No. 2. pp. 54-58.
8. Goroshko I.V. Matematicheskoe modelirovanie v upravlenii organami vnutrennih del. – M.: Akademiya upravleniya MVD Rossii, 2000. – 145 p.
9. Demidov Y.N. O tekushchih zadachah i problemnykh aspektah v sfere organizatsii deyatel'nosti podrazdelenij politsii po ohrane obshchestvennogo poryadka territorial'nykh organov MVD Rossii. Nauchnyj portal MVD Rossii, 2013, No. 4. - pp.104-110.
10. Klushin O.Z. Operativnaya obstanovka: ponyatie, analiz, prognoz. – M.: Akademiya upravleniya, 2010. – 256 p.
11. Nosko V.P. Ekonometrika dlya nachinayushchih. - M.: IEHPP, 2005. – 379 p.