СПОСОБ УТОЧНЕНИЯ РЕЗУЛЬТАТОВ ИНКЛИНОМЕТРИИ СКВАЖИН С ПОМОЩЬЮ РЕГРЕССИИ МЕТОДОМ ОПОРНЫХ ВЕКТОРОВ С БАЙЕСОВСКОЙ ОПТИМИЗАЦИЕЙ ПАРАМЕТРОВ

Рябов Антон Владимирович
кандидат техн. наук, доцент
Арзамасский политехнический институт (филиал) НГТУ
г. Арзамас

SUPPORT VECTOR REGRESSION WITH BAYESIAN PARAMETERS OPTIMIZATION FOR INCLINOMETER WELLS MEASUREMENTS REFINEMENT

Ryabov Anton Vladimirovich
Candidat of Science, assistant professor
of Arzamas Polytechnic Institute (branch) of N. Novgorod State Technic University,
Arzamas

АННОТАЦИЯ

В статье предложен способ уточнения результатов инклинометрии скважин с помощью регрессии данных показаний прибора методом опорных векторов с байесовской оптимизацией параметров алгоритма. Описан алгоритм, и представлен результат моделирования.

ABSTRACT

This paper propose a method of inclinometer wells measurements refinement via support vector regression with Bayesian parameters optimization. The results demonstrate that the given method-based algorithm is efficient for wells inclinometer measurements correction.

Ключевые слова: инклинометрия скважин; метод опорных векторов; регрессия; байесовская оптимизация.

Keywords: wells inclinometry; support vector machine; data regression; Bayesian optimization.

В процессе инклинометрии скважин неизбежны ошибки показаний прибора связанные с погрешностью датчиков информации, неравномерностью движения прибора, перекоса его в стволе скважины, ударами о выступы в породе, и т.д. В работе [4] предлагается для формирования траектории скважины использовать данные зенитного угла из измерений при подъеме прибора, а азимутальные данные – при его спуске, с последующим применением алгоритма сглаживания для устранения аномальных выбросов в измерениях, влияющих на точность траектории.

Целью настоящей работы является уточнение координат траектории скважины путем построения аппроксимирующей кривой методом опорных векторов по первичным данным инклинометрии, которая бы обеспечивала нечувствительность к выбросам в данных.

Метод опорных векторов (Support Vector Machin - SVM) - это алгоритм машинного обучения, представленный Вапником в 1995 году для задач бинарной классификации [1]. Затем результат был распространен и на задачи регрессионного анализа (Support Vector Regression - SVR). Цель метода - уместить как можно больше точек данных на полосе шириной ε. Для решения задач нелинейной регрессии применяют ядерные функции, этот прием известен как трюк с ядром (kernel trick), где предполагается, что существует некоторое спрямляющее пространство H, большей размерности, при переходе в которое нелинейная регрессия сводится к решению задачи линейной регрессии [3]. Данный метод уже достаточно хорошо описан, широко имплементирован, и зарекомендовал себя как мощный инструмент, однако по-прежнему мало используется из-за неопределенности параметров SVM-модели.

Запишем уравнение регрессии:

$$y = f(x) = \omega^T \phi(x) + b, \quad (1)$$

где $\phi(x): R \to F$ нелинейное отображение x в пространство более высокой размерности, $\omega \in F$ - вектор весовых коэффициентов, b - некоторая постоянная. Значения ω и b находим, решая задачу оптимизации, минимизирующую ω:

$$min\frac{1}{2}\|\omega\|^2, \text{ так чтобы}$$
$$\begin{cases} y_i = \omega^T \phi(x_i) - b \le \varepsilon, \\ y_i = \omega^T \phi(x_i) - b \ge -\varepsilon. \end{cases} \quad (2)$$

Используем фиктивные переменные ослабления ξ и ξ^* для штрафа функции потерь за выход из диапазона ε.

$$min\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \text{ при ограничениях}$$

$$\begin{cases} y_i = \omega^T \phi(x_i) - b \le \varepsilon + \xi_i, \\ y_i = \omega^T \phi(x_i) - b \ge -\varepsilon - \xi_i^*, \\ \xi_i, \xi_i^* \ge 0, i = 1,2, \ldots n. \end{cases} \quad (3)$$

Где C это постоянный положительный параметр регуляризации, ε - коэффициент нечувствительности, ξ и ξ^* - переменные ослабления, определяющие верхнее и нижнее отклонения соответственно, и n - это количество входных данных (Рис. 1).

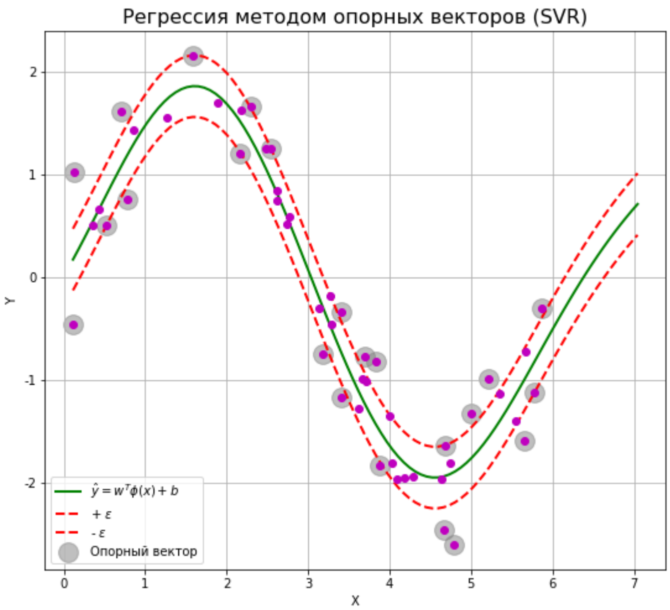

Рисунок 1. Иллюстрация метода опорных векторов

Используя Лагранжиан и соответствующие условия оптимальности, полученное уравнение перепишется как [2, 1]:

$$f(x) = \sum_{i=1}^{n}(\beta_i - \beta_i^*)K(x, x_i) + b, \quad (4)$$

где $\beta_i \neq 0$ и $\beta_i^* \neq 0$- множители Лагранжа, а $K(x, x_i)$ - функция ядра. Разности $\beta_i - \beta_i^*$ представляют собой весовые коэффициенты ν_i, характеризующие влияние соответствующих опорных векторов на формирование функции регрессии.

Существует ряд хорошо описанных ядерных функций для нелинейной регрессии. Одним из самых широко применяемых ядер является ядро на основе радиальной базисной функции (Radial Basis Function - RBF), или гауссово ядро:

$$K(x, x_i) = exp\left\{-\frac{\|x - x_i\|^2}{2\sigma^2}\right\}. \quad (5)$$

Это выражение часто упрощается до:

$$K(x, x_i) = exp(-\gamma\|x - x_i\|^2). \quad (6)$$

Где $\gamma = \frac{1}{2\sigma^2}$ - свободный параметр, который можно считать параметром отсечения для гауссовой сферы, или шириной ядра. В дальнейшем он потребует оптимизации.

Моделирование проводилось на массиве данных зенитного угла реальной траектории скважины. Значение зенитного угла при спуске и подъеме прибора на одной и той же глубине примем равными, и данные значения будем считать истинными. За «рабочие» значения показаний инклинометра примем шум вокруг истинных показаний прибора на спуске и подъеме (Рис. 2).

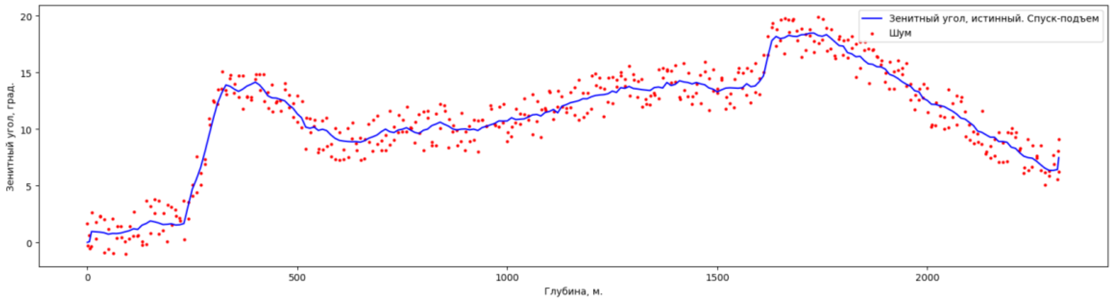

Рисунок 2. Зенитный угол и шум

Параметры SVR - C, γ и ϵ задаются экспертом [3], и заранее неизвестны. Произвольное задание этих параметров чаще всего вызывает негативные явления, именуемые переобучением и недообучением (Рис. 3). Имея диапазоны возможных значений этих параметров, необходимо отыскать их оптимальную комбинацию, обеспечивающую требуемую точность алгоритма. Одними из наиболее распространенных методов выбора параметров для SVM-алгоритмов являются прямые и кросс-валидационные поиски: по сетке (Grid search), и случайный поиск (Random search) (Рис. 4). Данные методы вычислительно дороги, поскольку осуществляют последовательный перебор всех возможных сочетаний параметров модели.

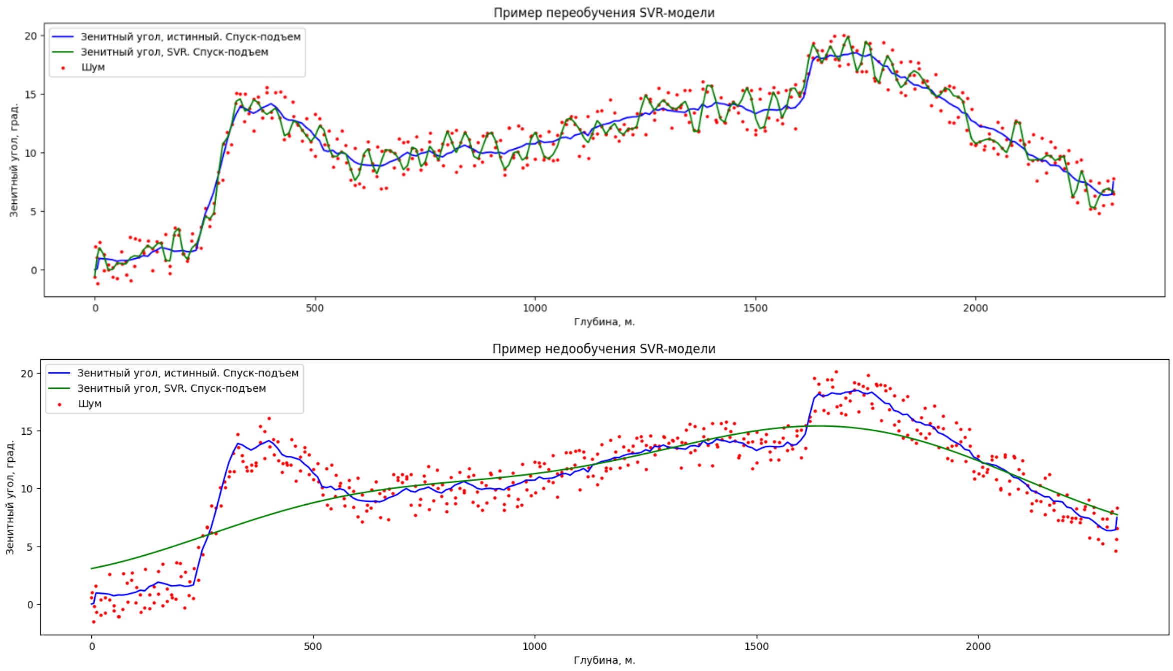

Рисунок 3. Пример переобучения и недообучения SVR-модели

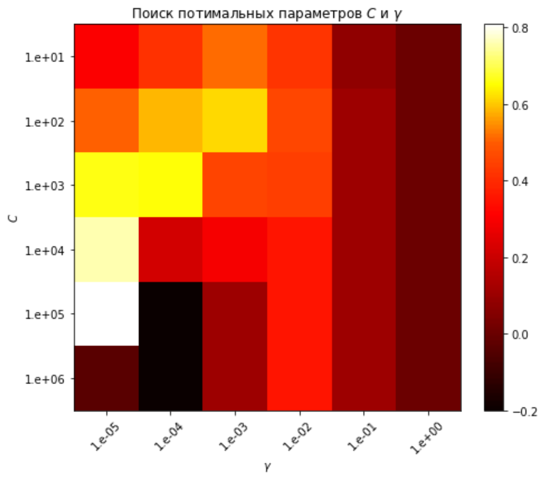

Рисунок 4. Поиск параметров по сетке значений

Байесовская оптимизация параметров SVM-регрессии

С ростом производительности вычислительной техники стали вновь актуальными методы оптимизации, в основе которых лежит байесовский подход. Байесовская оптимизация принимает во внимание оценки, полученные на предыдущих итерациях при выборе набора параметров для следующей оценки. Осознанно выбирая комбинации параметров, алгоритм сосредотачивается на тех областях пространства параметров, которые, по его мнению, принесут более высокие оценки валидации. Такой подход обычно требует меньшее количество итераций для получения оптимального набора значений параметров, поскольку игнорируются области пространства параметров, которые, по его мнению, не принесут улучшения. Байесовская оптимизация использует распределения вероятностей для каждого параметра, из которого будет производиться выборка. Эти распределения задаются пользователем. Задание распределения для каждого параметра - одна из необходимых частей процесса. Один из подходов здесь - начать с широкого диапазона значений параметров, и позволить алгоритму оптимизации самостоятельно определить область оптимальных значений. Затем в следующем прогоне можно сосредоточиться на некоторой окрестности найденной области для уточнения значений параметров.

Целевую функцию для оценки комбинаций параметров, и оптимальные параметры можно записать так:

$$\text{Точность} = f(C, \gamma).$$
$$\text{Параметры}_{\text{опт}} = argmax\, f\,(C, \gamma) \tag{7}$$

Обращение к целевой функции вычислительно дорого, поэтому необходима дополнительная функция, которая бы отслеживала текущие параметры, и давала прогноз целесообразности применения их для вычисления точности. Такой прогноз дает комбинация суррогатной функции и функция выбора. Их задача - предложить такие параметры, которые обеспечат наивысшую точность целевой функции. Суррогатная функция это приложение теоремы Байеса, по которой априорное распределение становится апостериорным каждый раз, когда появляется все большее количество данных (8). Данные в этом случае - это оценки точности, которые мы получаем от целевой функции по переданным ей параметрам.

$$P(\text{Точность}|\text{Параметры}),$$
$$P(y|x) = \frac{p(x|y)p(y)}{p(x)}. \tag{8}$$

Вероятность набора параметров с учетом точности $p(x|y)$, где x - набор параметров Cи γ, делится на два отдельных распределения: $l(x)$ и $g(x)$. Первое определяет распределение параметров, когда значение полученной отрицательной оценки точности ниже порогового значения оценки y^* (к примеру, y^* - максимальная оценка точности, известная на данный момент). Второе - распределение параметров для оценок выше этого же порога. $l(x)$ и $g(x)$ являются смесью гауссовых распределений, или линейной комбинацией распределений. Каждый раз, когда набор параметров оценивается целевой функцией, он попадает в соответствующее распределение $l(x)$ или $g(x)$ в зависимости от полученной оценки точности. Каждое распределение обновляется на следующей итерации, после чего оно используется для выборки следующей комбинации параметров, которая максимизирует ожидаемое улучшение.

$$P(\text{Параметры}|\text{Отрицательная оценка}) = \begin{cases} l(x), \text{ если отрицательная оценка} \prec y^* \\ g(x), \text{ если отрицательная оценка} \geq y^*. \end{cases} \tag{9}$$

В байесовской оптимизации параметры, предлагаемые для оценки целевой функцией, выбираются путем применения критерия к суррогатной функции. Этот критерий определяется функцией выбора. Распространенным подходом является использование показателя под названием «Ожидаемое улучшение» (Expected Improvement - EL). Для суррогатной функции ожидаемое улучшение можно определить следующим уравнением:

$$EL_{y^*}(x) = \frac{Yy^*l(x) - l(x)\int_{-\infty}^{y^*} p(y)dy}{Yl(x) + (1-Y)g(x)} \propto \left(Y + \frac{g(x)}{l(x)}(1-Y)\right)^{-1}. \tag{10}$$

Из этого уравнения видно - чтобы получить максимальное ожидаемое улучшение, необходимо максимизировать отношение $g(x)/l(x)$. Другими словами, необходима высокая вероятность набора параметров при отрицательной оценке точности ниже порога y^*. Алгоритм выбирает параметры из распределения $l(x)$, проверяет их в соотношении $g(x)/l(x)$, и предлагает набор, который соответствует наибольшему ожидаемому улучшению. Пороговое значение y^* определяется переменной Y, которая представляет собой квантиль отрицательных оценок точности, наблюдаемых до текущей итерации, для использования в качестве точки отсечения.

После установки начальных значений распределений параметров, алгоритм байесовской оптимизации выполняет следующие шаги:

Шаг 1. Получить набор параметров, которые максимизируют ожидаемое улучшение, оптимизируя функцию выбора суррогатной функцией.

Шаг 2. Используя этот набор параметров получить значение целевой функции.

Шаг 3. Используя формулу Байеса обновить суррогатную функцию значением целевой функции.

Шаг 4. Повторять шаги 1-3 пока не достигнуто заданное количество итераций.

На первых итерациях суррогатная функция показывает слабую аппроксимацию целевой функции, поскольку велика дисперсия исходных значений параметров. Но с увеличением числа итераций, оценки, полученные от целевой функции, используются для обновления распределений $l(x)$ и $g(x)$ с помощью байесовского выражения, которые становятся все более точным отражением фактического распределения комбинаций параметров.

На рис. 5 точками отмечены участки областей пространства параметров, исследованных алгоритмом на каждой итерации его выполнения. В данном эксперименте поиск оптимальных значений параметров проводился за 1500 итераций, но уже к трехсотой итерации видно, что алгоритм определил область оптимума, и далее продолжалось лишь уточнение.

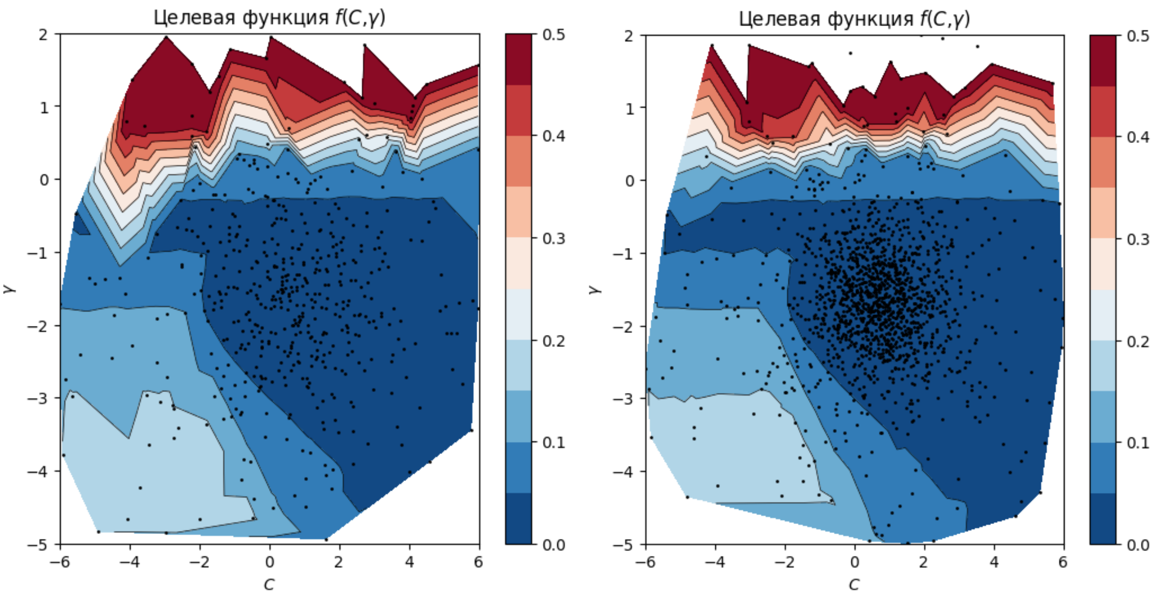

Рисунок 5. Поиск области оптимальных параметров. 300 и 1500 итераций.

Время выполнения алгоритмом поиска можно существенно сократить, введя контроль значений целевой функции на определенном количестве последних итераций, и останавливать работу при достижении дисперсии этих величин некоторого заданного значения.

Применение уточненных параметров к методу опорных векторов дает следующий результат (Рис. 6). Используя алгоритм для вектора значений азимута, и переведя данные в прямоугольную систему координат, можем сравнить траектории скважин – истинную, рабочую (шум), и скорректированную при помощи SVM-регрессии (Рис. 7). На первом фрагменте представлена скважина полностью, на втором – 100 метров до входа в круг допуска радиусом 30 метров.

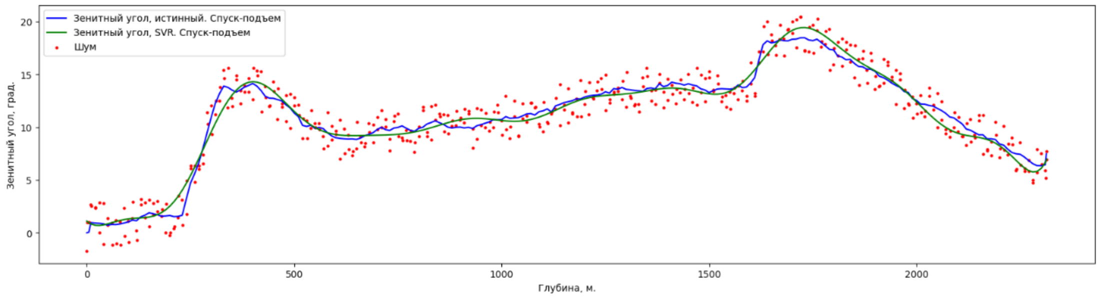

Рисунок 6. Регрессия данных методом опорных векторов

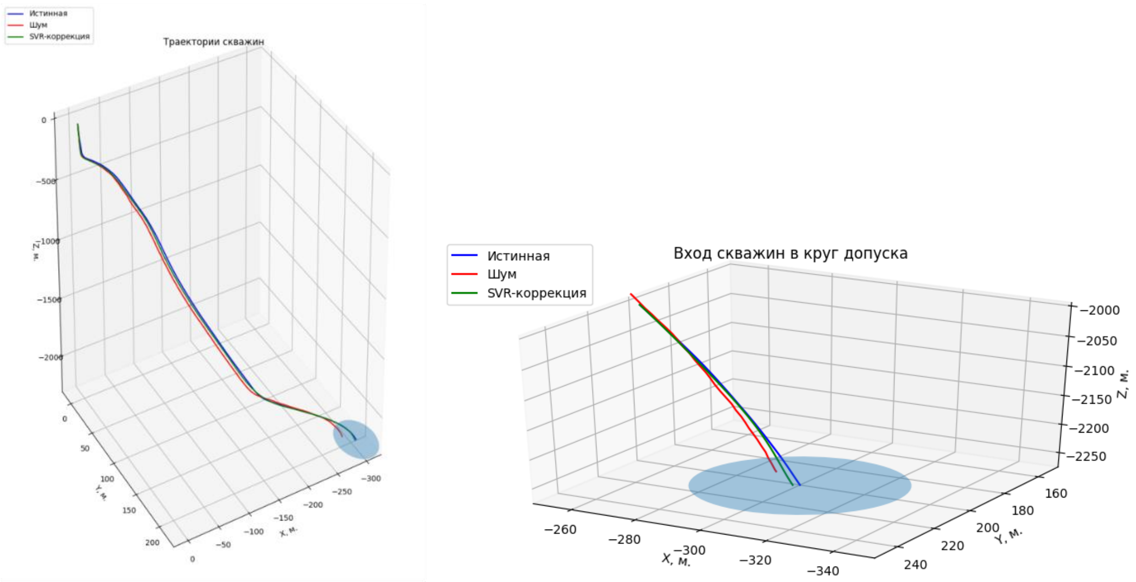

Рисунок 7. Траектории скважин

Представленный способ уточнения данных инклинометрии скважин был протестирован на нескольких траекториях, полученных прибором ИГН73-100/80, и показал возможность коррекции координат точки входа в круг допуска на 5-30 метров при различных уровнях шума.

Литература

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and computing 14, 199–222.

Yeh, C.Y., Huang, C.W., Lee, S.J., 2011. A multiple-kernel support vector regression approach for stock market price forecasting. Expert Systems with Applications 38, 2177–2186.

М.А. Борисов, А.А. Гуськов, С.И. Кошелев, 2016. Повышение точности определения траектории скважины гироинклинометром за счет вторичной обработки данных. Приволжский научный вестник № 12-2 (64) – 2016, 11-14.

ВЕРОЯТНОСТНЫЙ ПОДХОД К ОЦЕНКЕ ЗАЩИЩЁННОСТИ ИНФОРМАЦИИ ОТ УГРОЗ

Титов Михаил Юрьевич
кандидат тех. наук, с.н.с
РТУ МИРЭА
г. Москва
Карасев Павел Игоревич
кандидат тех. наук, доцент
РТУ МИРЭА
г. Москва
Пушкин Павел Юрьевич
кандидат тех. наук, доцент
РТУ МИРЭА
г. Москва
Титова Маргарита Михайловна
инженер-конструктор первой категории
ООО «Прогресстех»
г. Москва

PROBABLE APPROACH TO SAFETY ASSESSMENT INFORMATION FROM THREATS

Titov Michail
Candidate of Science, assistant professor
of RTU MIREA, Moscow
Karasev Pavel
Candidate of Science, assistant professor
of RTU MIREA, Moscow
Pushkin Pavel
Candidate of Science, assistant professor
of RTU MIREA, Moscow
Titova Margarita
design engineer of the first category
LLC "Progresstech"

DOI: 10.31618/NAS.2413-5291.2021.1.74.515

АННОТАЦИЯ

Общеизвестно, что основная цель защиты информации – обеспечение заданного уровня её безопасности. Заданный уровень безопасности информации характеризуется состоянием её защищённости от угроз, при котором обеспечивается допустимый риск её уничтожения, изменения, хищения, а также блокирования.

Риски зависят от уровня инженерно-технической защиты информации (ИТЗИ), который определяется ресурсами системы. Чем больше ресурс на защиту информации, тем выше уровень безопасности. При неограниченном ресурсе можно получить сколь угодно малую вероятность реализации угрозы.

ABSTRACT

It is well known that the main goal of information protection is to ensure a given level of its security. The specified level of information security is characterized by the state of its protection from threats, which provides an acceptable risk of its destruction, alteration, theft, and blocking.

Risks depend on the level of engineering and technical protection of information