УДК 519.862.6
DOI 10.52575/2687-0932-2022-49-1-121-133

Программный комплекс построения квазилинейных регрессий по критериям точности и нелинейности

Караулова А.В., Базилевский М.П.
Иркутский государственный университет путей сообщения,
Россия, 664074, Иркутская обл., г. Иркутск, ул. Чернышевского, 15
E-mail: anuta_kav@mail.ru, mik2178@yandex.ru

Аннотация. Работа посвящена проблеме выбора структурных спецификаций квазилинейных регрессионных моделей. Такие регрессии довольно просто оцениваются, но наличие в них нелинейных преобразований переменных затрудняет интерпретацию их оценок. Ранее авторами была сформулирована двухкритериальная задача выбора спецификации квазилинейной регрессии, состоящая в максимизации коэффициента детерминации и одновременной минимизации общего критерия нелинейности. Для большого числа переменных такая формулировка превращается в сложную вычислительную задачу. В настоящей работе впервые описан разработанный нами программный комплекс КВАРТОН-1, полностью автоматизирующий решение двухкритериальной задачи. В нем предусмотрена работа в двух режимах. В первом из них формируется множество Парето, с помощью которого пользователь может визуально прослеживать последовательную трансформацию линейной регрессии в нелинейную модель и выбирать наиболее приемлемую альтернативу. Во втором режиме сначала автоматически исключаются все значительно нелинейные регрессии, а из оставшихся выбирается лучшая по критерию детерминации. В обоих режимах предусмотрена возможность исключения моделей с незначимыми по t-критерию Стьюдента коэффициентами. С помощью КВАРТОН-1 была успешно решена задача моделирования грузовых железнодорожных перевозок в Иркутской области.

Ключевые слова: регрессионная модель, квазилинейная регрессия, коэффициент детерминации, критерий нелинейности, t-критерий Стьюдента, интерпретация, грузовые железнодорожные перевозки

Для цитирования: Караулова А.В., Базилевский М.П. 2022. Программный комплекс построения квазилинейных регрессий по критериям точности и нелинейности. Экономика. Информатика, 49(1): 121–133. DOI 10.52575/2687-0932-2022-49-1-121-133

Software complex for constructing quasi-linear regressions according to the criteria of accuracy and non-linearity

Anna V. Karaulova, Mikhail P. Bazilevskiy
Irkutsk State Transport University
15 Chernyshevskogo St, Irkutsk, 664074, Russia
E-mail: anuta_kav@mail.ru, mik2178@yandex.ru

Abstract. This article is devoted to the problem of choosing structural specifications for quasi-linear regression models. Such regressions are fairly easy to estimate, but the presence of non-linear transformations of variables in them makes it difficult to interpret their estimates. Previously, the authors formulated a two-criterion problem of choosing the specification for quasi-linear regression, which consists in maximizing the coefficient of determination and simultaneously minimizing the general criterion of nonlinearity. For a large number of variables, such a formulation turns into a complex computational problem. In this paper, we describe for the first time a software complex developed by us that fully automates the solution of a two-criteria problem. It provides work in two modes. In the first of them, a Pareto set is formed, with the help of which the user can visually trace the sequential transformation of a linear

regression into a nonlinear model and choose the most acceptable alternative. In the second mode, all significantly non-linear regressions are first automatically excluded, and the best one according to the determination criterion is selected from the remaining ones. In both modes, it is possible to exclude models with coefficients that are insignificant according to Student's t-test. With the help of the software complex, the problem of modeling freight rail transportation in the Irkutsk region was successfully solved.

Keywords: regression model, quasi-linear regression, coefficient of determination, criterion of non-linearity, Student's t-test, interpretation, freight rail transportation

For citation: Karaulova A.V., Bazilevskiy M.P. 2022. Software complex for constructing quasi-linear regressions according to the criteria of accuracy and non-linearity. Economics. Information technologies, 49(1): 121–133 (in Russian). DOI 10.52575/2687-0932-2022-49-1-121-133

Введение

При проведении регрессионного анализа [Brook, Arnold, 2018; Lawrence, 2019; Montgomery et al., 2021] исследователи, как правило, отдают предпочтение построению линейных моделей, которые легко оцениваются, а их оценки отчетливо интерпретируются. Однако реальные социально-экономические процессы редко носят абсолютно линейный характер, поэтому для их адекватного описания приходится прибегать к построению нелинейных зависимостей [Mize, 2019; Borowiak, 2020]. Различают два класса нелинейных регрессий. К первому классу относят нелинейные по параметрам модели, для оценивания которых приходится прибегать к реализации специальных численных методов. Ко второму классу относят квазилинейные модели (нелинейные по переменным, но линейные по параметрам), которые тем или иным способом могут быть линеаризованы.

Очевидно, что структурных спецификаций квазилинейных регрессий существует бесчисленное множество. Проблема в том, как для имеющихся в наличии статистических данных выбрать из всего этого многообразия наиболее адекватный вариант зависимости. В некоторых случаях вид спецификации устанавливается исходя из экономического смысла рассматриваемых переменных. Иногда применяется графический метод, когда по точечным диаграммам выбираются подходящие преобразования переменных. Самым качественным, но трудоемким методом решения проблемы спецификации является перебор всех возможных регрессий и выбор лучшей из них на основе некоторого критерия качества. В [Носков, 1996; Носков, 2021б] для выбора спецификации предложена технология организации «конкурса» моделей. В [Базилевский, Носков, 2012] приводится описание программного комплекса автоматизации процесса построения регрессионных моделей (ПК АППРМ), предназначенного для проведения «конкурса» моделей. С помощью ПК АППРМ решено множество прикладных задач анализа данных (см., например, [Носков, Врублевский, 2016; Баенхаева и др., 2016]). Полученные в результате организации «конкурса» квазилинейные регрессии зачастую обладают высоким аппроксимационным качеством, поэтому могут быть использованы для получения прогнозных значений зависимой переменной. Но при этом такие модели получаются значительно нелинейными, поэтому возникают трудности с интерпретацией их оценок.

В рамках регрессионного анализа разработан целый арсенал различных критериев адекватности [Cavanaugh, Neath, 2019; Chicco et al., 2021; Носков, 2021а]. Для количественной оценки степени нелинейности квазилинейных регрессий в [Базилевский, 2018] были предложены специальные критерии. В [Базилевский, Караулова, 2021] на основе этих критериев была сформулирована двухкритериальная задача, состоящая в максимизации коэффициента детерминации и одновременной минимизации критерия нелинейности. Целью настоящей работы является описание разработанного нами для решения сформулированной двухкритериальной задачи программного комплекса и демонстрация его работы на примере решения конкретной прикладной задачи анализа данных.

Двухкритериальная задача

Квазилинейная регрессионная модель имеет вид:

$$y_i = \alpha_0 + \sum_{k=1}^{q}\sum_{j=1}^{m}\alpha_{kj} f_k(x_{ij}) + \varepsilon_i, \qquad i=\overline{1,n}, \tag{1}$$

где n – объем выборки; m – число объясняющих переменных; q – число элементарных функций; y_i, $i=\overline{1,n}$ – значения объясняемой переменной; $x_{ij}>0$, $i=\overline{1,n}$, $j=\overline{1,m}$ – значения объясняющих переменных; $f_k(x)$, $k=\overline{1,q}$ – элементарные функции; α_0, α_{kj}, $k=\overline{1,q}$, $j=\overline{1,m}$ – неизвестные параметры; ε_i, $i=\overline{1,n}$ – ошибки аппроксимации.

Будем считать, что входящие в модель (1) элементарные функции являются непрерывными и монотонными на отрезках $[x^j_{\min}, x^j_{\max}]$, $j=\overline{1,m}$, где $x^j_{\min}=\min\{x_{1j}, x_{2j}, ..., x_{nj}\}$, $x^j_{\max}=\max\{x_{1j}, x_{2j}, ..., x_{nj}\}$, $j=\overline{1,m}$.

Пусть в модель (1) каждая преобразованная объясняющая переменная входит ровно 1 раз, тогда её можно представить в виде:

$$y_i = \alpha_0 + \sum_{j=1}^{m}\alpha_j f_{\omega_j}(x_{ij}) + \varepsilon_i, \qquad i=\overline{1,n}, \tag{2}$$

где $\omega_j \in [1,q]$, $j=\overline{1,m}$ – элементы индексного вектора Ω, каждый из которых показывает номер элементарной функции для преобразования j-й объясняющей переменной.

Известно, что если регрессионная модель (2) оценивается с помощью метода наименьших квадратов (МНК), то об её адекватности можно судить по значению коэффициента детерминации, который находится по формуле:

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}},$$

где RSS – сумма квадратов остатков регрессии; TSS – общая сумма квадратов.

Коэффициент детерминации R^2 принимает значения от 0 до 1. Чем ближе его значение к 1, тем сильнее зависимость.

Для оценки степени нелинейности преобразованных переменных в квазилинейных регрессиях (2) в [Базилевский, 2018] разработаны критерии нелинейности «по площади» NC^j_S, $j=\overline{1,m}$, которые находятся по формулам:

$$NC^j_S = \left| \frac{f_{\omega_j}(x^j_{\max}) + f_{\omega_j}(x^j_{\min})}{f_{\omega_j}(x^j_{\max}) - f_{\omega_j}(x^j_{\min})} - \frac{2\int_{x^j_{\min}}^{x^j_{\max}} f_{\omega_j}(x_j)dx_j}{\left(x^j_{\max} - x^j_{\min}\right)\left(f_{\omega_j}(x^j_{\max}) - f_{\omega_j}(x^j_{\min})\right)} \right|, \qquad j=\overline{1,m}. \tag{3}$$

Область значений каждого из критериев (3) $NC^j_S \in [0,1]$. Если $NC^j_S = 0$, то преобразование j-й объясняющей переменной является линейным, а если $NC^j_S \to 1$, то оно в значительной степени нелинейно. Если значение NC^j_S близко к нулю, то вместо оцененного коэффициента $\tilde{\alpha}_j$ модели (2) можно интерпретировать коэффициент

$$k_j = \tilde{\alpha}_j \frac{f_{\omega_j}(x^j_{\max}) - f_{\omega_j}(x^j_{\min})}{x^j_{\max} - x^j_{\min}}. \tag{4}$$

Для оценки степени нелинейности квазилинейной регрессии (2) в целом применяется критерий верхней границы нелинейности:

$$L = \max\{NC^1_S, NC^2_S, ..., NC^m_S\}.$$

Этот критерий обладает теми же свойствами, что и его компоненты.

В работе [Базилевский, Караулова, 2021] сформулирована двухкритериальная задача: требуется выбрать такие компоненты вектора Ω в регрессионной модели (2), которые обеспечивали бы как ее высокое аппроксимационное качество, так и низкую степень нелинейности, т. е.

$$R^2 \to \max, \; L \to \min. \tag{5}$$

В [Базилевский, Караулова, 2021] предложено 2 следующих способа решения задачи (5).

1. Сначала оценить все возможные спецификации модели (2). Затем сформировать множество Парето, по которому путем визуального просмотра значений критериев выбрать приемлемую альтернативу.

2. Сначала исключить все преобразованные переменные, для которых $NC_S^j > \delta$, где δ – некоторое число из промежутка $[0{,}1; 0{,}3]$. Затем из оставшихся переменных сформировать и оценить все возможные спецификации модели (2) и выбрать лучшую по величине R^2.

Первый способ требует оценки всех q^m регрессионных моделей, а второй – $\prod_{j=1}^{m} q_j$, где q_j – количество преобразований j-й объясняющей переменной, для которых $NC_S^j \le \delta$.

Программный комплекс

Для решения двухкритериальной задачи (5) на языке программирования Python был разработан программный комплекс построения квазилинейных регрессий по критериям точности и нелинейности «по площади» (ПК КВАРТОН-1). Поскольку решение задачи (5) может быть найдено двумя способами, то в ПК КВАРТОН-1 реализовано два соответствующих режима работы. При этом в обоих режимах предусмотрена возможность исключения моделей с незначимыми по t-критерию Стьюдента коэффициентами. В результате работы программного комплекса в зависимости от выбранного режима автоматически формируется отчет в Microsoft Word, содержащий полную информацию о построенных моделях и их характеристиках. ПК КВАРТОН-1 является универсальным программным продуктом, позволяющим решать задачи анализа данных из самых разных предметных областей.

Блок-схема алгоритма работы ПК КВАРТОН-1 представлена на рисунке 1. Кратко охарактеризуем каждый из этапов.

Сначала пользователь должен задать следующие параметры.

1. Имя файла со статистическими данными. Этот файл должен иметь расширение *.txt и располагаться в папке с программой. К содержимому файла предъявляются следующие требования: в первой строке содержатся имена переменных, отделенные друг от друга символом «табуляция»; первой в списке переменных указана зависимая переменная; в последующих строках значения переменных также отделяются друг от друга символом «табуляция»; разделителем целых и дробных частей вещественных чисел является точка.

2. Вектор преобразований независимых переменных transf. На данный момент в ПК КВАРТОН-1 реализовано 10 таких преобразований: x, x^2, x^3, $x^{-0{,}5}$, x^{-1}, x^{-2}, x^{-3}, $\sqrt{x}$, $\ln x$, 2^{ax}, где a – заданное число. Каждое из них зашифровано числами от 0 до 9. Например, если transf = [0,4], то в качестве преобразований выступают x и x^{-1}.

3. Значение переменной Rezh, отвечающей за режим работы программы. Если Rezh = 1, то система работает в режиме № 1, согласно которому осуществляется формирование множества всех возможных вариантов моделей, их оценивание с помощью МНК и формирование множества Парето на основе критериев R^2 и L. Если Rezh = 2, то система работает в режиме № 2, в котором на первом этапе отсекаются все сильно нелинейные преобразованные переменные, а уже потом из оставшихся формируется и оценивается множество возможных вариантов моделей, из которого выбирается единственная лучшая по критерию R^2 регрессия.

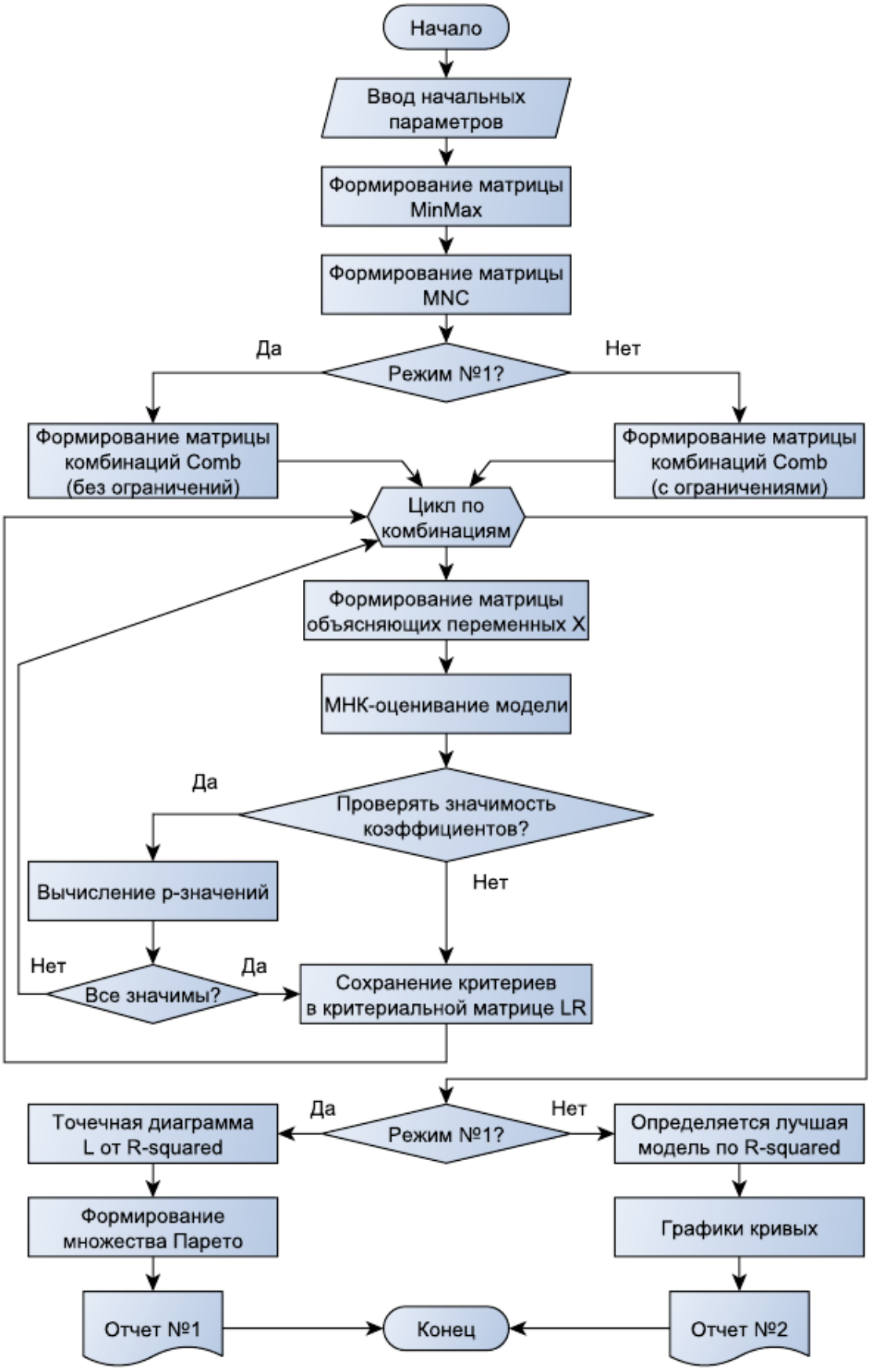

Рис. 1. Блок-схема алгоритма работы программы
Fig. 1. Block diagram of the program operation algorithm

Границу нелинейности Nelin (для режима № 2). По умолчанию Nelin = 0,2. Преобразованные переменные, у которых значение критерия нелинейности «по площади» будет превосходить величину Nelin, будут отсеяны.

Значение переменной Znach, отвечающей за необходимость отсечения в обоих режимах моделей, в которых хотя бы один коэффициент незначим по t-критерию Стьюдента на уровне Ur. Если Znach = 1, то исключение установлено, а если Znach = 2, то нет.

4. Уровень значимости Ur (для исключения моделей с незначимыми по t-критерию Стьюдента коэффициентами). По умолчанию Ur = 0,1.

После ввода начальных параметров и запуска ПК КВАРТОН-1 первым делом формируется матрица MinMax размера $2\times m$, первая строка которой содержит минимальные значения каждой объясняющей переменой, а вторая – их максимальные значения.

Затем на основе матрицы MinMax формируется матрица критериев нелинейности переменных MNC размера $m\times q$. На пересечении i-й строки и j-го столбца этой матрицы содержится значение критерия нелинейности i-й переменной, преобразованной с помощью j-й функции. Элементы матрицы MNC определяются по формуле (3).

Далее в зависимости от выбранного режима формируется матрица комбинаций моделей Comb. Для этого использован модуль itertools языка Python. В режиме № 1 нет ограничений на значения критериев нелинейности «по площади», поэтому формируются все возможные комбинации преобразованных переменных, общее число которых есть число размещений с повторениями из q элементов вектора transf по m. В этом случае матрица Comb имеет размер $q^m\times m$.

В режиме № 2 сначала исключаются те преобразованные переменные, для которых значения критериев нелинейности превышают величину Nelin. В результате для каждой объясняющей переменной из набора x_1, x_2, …, x_m формируется свой вектор преобразованных переменных transf_1, transf_2, …, transf_m, состоящий из q_1, q_2, …, q_m преобразований соответственно. Если все эти векторы не являются пустыми, то матрица Comb находится как их декартово произведение и имеет размер $\left(\prod_{j=1}^{m} q_j\right)\times m$.

После формирования матрицы Comb запускается общий для обоих режимов цикл по всем её строкам. Внутри него сначала для каждой комбинации преобразований формируется матрица значений объясняющих переменных X. Затем с помощью модуля statsmodels языка Python на основе матрицы X находятся МНК-оценки модели. Далее, если пользователем не выбрана опция исключения моделей с незначимыми по t-критерию Стьюдента коэффициентами (Znach = 2), то характеристики регрессии (её номер, значение критериев L и R^2) сохраняются в очередной строке критериальной матрицы LR. Если опция исключения моделей выбрана (Znach = 1), то для всех коэффициентов определяются P-значения и проверяются условия их превышения заданному уровню значимости Ur. Если хотя бы одно условие выполняется, то характеристики регрессии не сохраняются в критериальной матрице LR. В противном случае, когда все коэффициенты значимы, информация о модели заносится в LR.

После завершения формирования критериальной матрицы LR алгоритм работы ПК КВАРТОН-1 разветвляется в зависимости от выбранного режима. В режиме № 1 на основе матрицы LR строится точечная диаграмма зависимости значений критерия L от R^2, а также формируется множество Парето для двухкритериальной задачи (5). Затем автоматически формируется отчет № 1 в Microsoft Word, содержащий исходные статистические данные, матрицу критериев нелинейности переменных MNC, параметры поиска (информация о выбранном режиме и опции исключения моделей), количество проанализированных моделей, точечную диаграмму зависимости L от R^2 и уравнения регрессий, входящих во множество Парето. Для каждого уравнения в этом множестве указаны значения критериев L и R^2, ин-

формация о значимости его коэффициентов в словесной форме, а также угловые коэффициенты (4), которые можно использовать для интерпретации вместо оцененных коэффициентов. В режиме № 2 на основе матрицы LR выбирается лучшая по величине коэффициента R^2 модель, для которой строятся графики кривых, отражающих нелинейность её переменных. Затем автоматически формируется отчет № 2, который отличается от отчета № 1 тем, что в нем вместо точечной диаграммы и множества Парето содержится уравнение одной единственной наилучшей регрессии и соответствующие ей графики кривых.

Необходимо выделить следующие две особенности ПК КВАРТОН-1.

1. Программный комплекс не предназначен для решения задач отбора наиболее «информативных» объясняющих переменных [Носков, 1996]. Иными словами, все объясняющие переменные, указанные в исходном наборе данных, включаются явно или в виде преобразований в каждую оцененную в результате работы программы спецификацию модели.

2. Желательно работать с программным комплексом после проведения процедуры отбора наиболее «информативных» переменных, заранее получив адекватную модель со всеми значимыми и удовлетворяющую содержательному смыслу задачи коэффициентами. Таким образом, ПК КВАРТОН-1 выполняет функцию улучшения выбранной на начальном этапе хорошей модели.

Решение прикладной задачи

Актуальной научной задачей является моделирование перевозочных процессов [Щербанин и др., 2017; Lee, Kim, 2018; Khan M., Khan F, 2020; Gao et al., 2020; Pompigna, Mauro, 2020]. Для демонстрации работы ПК КВАРТОН-1 решалась задача моделирования грузовых железнодорожных перевозок в Иркутской области. Для этого с использованием архивов Росстата (https://rosstat.gov.ru/) были собраны статистические данные (табл. 1) за период с 2000 г. по 2018 г. по следующим переменным:

y – отправление грузов железнодорожным транспортом общего пользования (млн тонн);

x_8 – число собственных легковых автомобилей на 1000 человек населения (штук);

x_{18} – число предприятий и организаций (тысяч);

x_{36} – удельный вес автомобильных дорог с усовершенствованным покрытием в протяженности автомобильных дорог с твердым покрытием общего пользования (в процентах);

x_{39} – среднегодовая номинальная начисленная заработная плата работников транспорта (тысяч рублей);

x_{58} – тарифы на грузовые перевозки (железнодорожный транспорт), найденные через соответствующие индексы при величине реального тарифа в 2001 г., взятого равным 1000 усл. ед.

Заметим, что изначальное количество объясняющих переменных было равно 62. После применения метода исключения их осталось 5.

С помощью МНК по исходным данным была оценена модель множественной линейной регрессии:

$$\tilde{y} = -54,059 - \underset{(-2,477)}{0,15}\, x_8 + \underset{(4,385)}{0,899}\, x_{18} + \underset{(1,873)}{2,394}\, x_{36} + \underset{(2,147)}{1,888}\, x_{39} - \underset{(-2,184)}{0,0216}\, x_{58}, \qquad (6)$$

для которой коэффициент детерминации $R^2 = 0,8158$. В скобках под коэффициентами регрессии указаны значения t-критерия Стьюдента. Как видно, все коэффициенты значимы и соответствуют экономическому смыслу задачи.

Затем проводилось моделирование с использованием ПК КВАРТОН-1. Для этого было использовано 9 элементарных функций: x, x^2, x^3, $x^{-0,5}$, $\frac{1}{x}$, $\frac{1}{x^2}$, $\frac{1}{x^3}$, $\sqrt{x}$, $\ln x$. Для построения множества Парето был выбран первый режим (Rezh = 1). И для отсечения уравнений с незначимыми коэффициентами было задано Znach = 1 для уровня значимости Ur = 0,1.

Таблица 1
Table 1

Статистические данные
Statistical data

Год	y	x_8	x_{18}	x_{36}	x_{39}	x_{58}
2000	50,7	138,8	46,041	39,3	3,537	656,102
2001	51,6	140	49,542	39,8	4,804	1000
2002	50,2	147,2	52,724	39,8	6,385	1203
2003	58,2	151,6	56,566	40	8,166	1283,601
2004	61,7	156,5	59,127	40,2	10,353	1450,469
2005	68,1	143,3	63,414	40,3	12,067	1624,525
2006	68,7	154,8	65,351	40,2	13,992	1746,852
2007	66,7	169,2	68,652	40,2	16,429	1895,334
2008	67,2	188,2	68,049	40,4	20,64	2354,763
2009	59,6	189,8	70,896	40,4	23,759	2613,081
2010	64,5	202,6	65,839	40,6	27,899	2859,494
2011	59,3	224,3	61,591	41,3	30,605	2999,324
2012	59,3	251,5	62,285	38,7	33,76	3088,404
2013	57,6	271,8	64,761	39,3	36,198	3269,075
2014	53,6	270,5	66,593	39,5	38,142	3361,59
2015	50,6	271,3	68,106	38,4	39,939	3799,605
2016	49,1	242,7	64,669	37,7	42,755	4092,555
2017	50,4	246,2	62,156	37,9	46,107	4244,389
2018	50	245,6	59,557	37,4	49,78	4472,312

В результате работы ПК КВАРТОН-1 был сформирован отчет в формате Microsoft Word. Приведем основное его содержимое.

В таблице 2 содержится матрица критериев нелинейности MNC.

Таблица 2
Table 2

Матрица критериев нелинейности (MNC)
Matrix of non-linearity criteria (MNC)

Преобр. / Перем.	x	x^2	x^3	$x^{-0,5}$	$\frac{1}{x}$	$\frac{1}{x^2}$	$\frac{1}{x^3}$	$\sqrt{x}$	$\ln x$
x_8	0	0,108	0,2086	0,1664	0,2207	0,3239	0,4173	0,0555	0,1112
x_{18}	0	0,0709	0,1396	0,1075	0,143	0,2126	0,2792	0,0358	0,0717
x_{36}	0	0,0165	0,033	0,0248	0,0331	0,0496	0,066	0,0083	0,0165
x_{39}	0	0,2891	0,4623	0,5791	0,7175	0,8673	0,9246	0,193	0,3966
x_{58}	0	0,248	0,4188	0,4461	0,5704	0,7441	0,8376	0,1487	0,3018

Как видно по таблице 2, самым нелинейным преобразованием оказалось x^{-3} для переменной x_{39}.

В ПК КВАРТОН-1 автоматически было оценено 59049 регрессионных моделей. Из них в 1396 регрессиях все коэффициенты оказались значимы по t-критерию Стьюдента для уровня значимости Ur = 0,1. Эти модели нанесены точками на точечную диаграмму, представленную на рисунке 2.

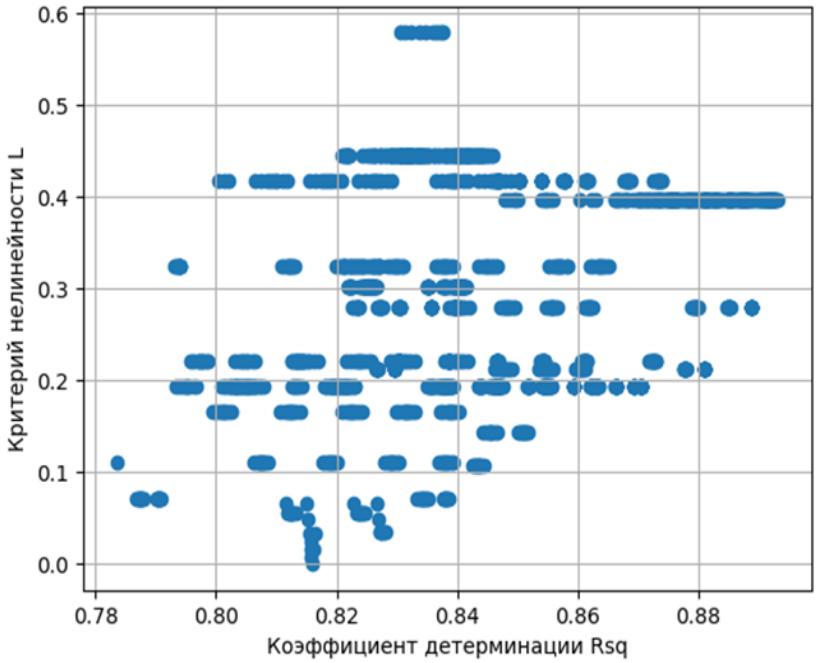

Рис. 2. Точечная диаграмма для критериев R^2 и L
Fig. 2. Scatter plot for R^2 and L criteria

По рисунку 2 видно, что для исходных данных существует довольно много качественных и вполне интерпретируемых моделей, для которых $R^2 \geq 0{,}8$, а $L \leq 0{,}2$.

Сформированное в ПК КВАРТОН-1 множество Парето представлено в таблице 3.

Таблица 3
Table 3

Множество Парето
Pareto set

№	Преобразованные переменные	L	R^2
1	2	3	4
1	$x_8, x_{18}, x_{36}, x_{39}, x_{58}$	0	0,8159
2	$x_8, x_{18}, x_{36}^2, x_{39}, x_{58}$	0,0165	0,8161
3	$x_8, x_{18}, x_{36}^3, x_{39}, x_{58}$	0,033	0,8163
4	$x_8, \sqrt{x_{18}}, x_{36}^3, x_{39}, x_{58}$	0,0358	0,828
5	$x_8, \ln x_{18}, x_{36}^3, x_{39}, x_{58}$	0,0717	0,8387
6	$\sqrt{x_8}, x_{18}^{-0,5}, x_{36}^3, x_{39}, x_{58}$	0,1075	0,8443
7	$\sqrt{x_8}, x_{18}^{-1}, x_{36}^3, x_{39}, x_{58}$	0,143	0,8519

Окончание табл. 3

1	2	3	4
8	$\sqrt{x_8}, x_{18}^{-1}, x_{36}^{-3}, \sqrt{x_{39}}, \sqrt{x_{58}}$	0,193	0,8706
9	$\ln x_8, x_{18}^{-2}, x_{36}^{3}, \sqrt{x_{39}}, \sqrt{x_{58}}$	0,2126	0,8811
10	$\ln x_8, x_{18}^{-3}, x_{36}^{3}, \sqrt{x_{39}}, \sqrt{x_{58}}$	0,2792	0,8888
11	$x_8^{-0,5}, x_{18}^{-3}, x_{36}^{-3}, \ln x_{39}, \ln x_{58}$	0,3966	0,8933

Как и следовало ожидать, нулевое значение критерия нелинейности L в таблице 3 получено для линейной регрессии (6). Вместе с тем она имеет и самое низкое значение коэффициента детерминации $R^2 = 0,8159$. Лучшей по величине коэффициента детерминации в таблице 3 оказалась модель № 11, для которой $R^2 = 0,8933$. Но она же является и самой нелинейной по величине L. Таким образом, с помощью таблицы 3 можно проследить плавную трансформацию линейной регрессии в нелинейную модель с наибольшим коэффициентом детерминации.

Уравнение регрессии № 11 в таблице 3 имеет вид:

$$\tilde{y} = \underset{}{262,786} + \underset{(3,185)}{826,754}\, x_8^{-0,5} - \underset{(-2,057)}{1771236,248} x_{18}^{-3} - \underset{(-2,163)}{1576200,523}\, x_{36}^{-3} + \underset{(3,758)}{34,794} \ln x_{39} - \underset{(-3,706)}{43,256} \ln x_{58}. \quad (7)$$

Все коэффициенты этого уравнения значимы.

Для модели (7) критерии нелинейности для преобразованных переменных x_{18}^{-3}, $\ln x_{39}$ и $\ln x_{58}$ превышают величину 0,2 и составляют соответственно 0,2792, 0,3966 и 0,3018. Поэтому коэффициенты при этих преобразованиях в уравнении (7) некорректно интерпретировать с помощью показателей (4). Стоит заметить, что переменные, преобразованные с помощью элементарной функции $\ln x$, всё же могут быть всегда интерпретированы по правилу: если x_j изменится на 1 %, то y изменится примерно на $\dfrac{\tilde{\alpha}_j}{100}$ единиц. Однако даже в этом случае не ясно, как объяснить преобразование x_{18}^{-3}.

Затем ПК КВАРТОН-1 был запущен во втором режиме (Rezh = 2) для тех же элементарных функций при Znach = 1, Ur = 0,2, Nelin = 0,2. В этом случае было автоматически проанализировано 1260 моделей, из которых в 429 регрессиях все коэффициенты оказались значимы по t-критерию Стьюдента. Лучшей из них оказалась регрессия № 8 из таблицы 3, имеющая уравнение:

$$\tilde{y} = \underset{}{209,482} - \underset{(-3,193)}{4,57} \sqrt{x_8} - \underset{(-3,968)}{2572,246}\, x_{18}^{-1} - \underset{(-1,849)}{1556602,187}\, x_{36}^{-3} + \underset{(3,104)}{18,477} \sqrt{x_{39}} - \underset{(-3,116)}{2,183} \sqrt{x_{58}}\ . \quad (8)$$

Все коэффициенты этого уравнения значимы.

Уравнение (8) является неким компромиссом между линейной регрессией (6) и сильно нелинейной, поэтому плохо интерпретируемой, регрессией (7).

Для модели (8) в ПК КВАРТОН-1 автоматически для каждой преобразованной переменной на интервале $[x_{\min}^j, x_{\max}^j]$ построена кривая $f_{\omega_j}(x_j)$, а через её концы проведена соответствующая прямая (рис. 3).

Также автоматически для интерпретации уравнения (8) были определены показатели (4): $k_{x_8} = -0,1616$, $k_{x_{18}} = 0,788$, $k_{x_{36}} = 1,9637$, $k_{x_{39}} = 2,0676$, $k_{x_{58}} = -0,0236$. Эти коэффициенты интерпретируется по такому же принципу, как и оценки в линейной регрессии (6). Например, в Иркутской области с ростом числа предприятий и организаций x_{18} на 1 тысячу отправле-

ние грузов железнодорожным транспортом общего пользования y увеличивается примерно на 0,788 млн тонн.

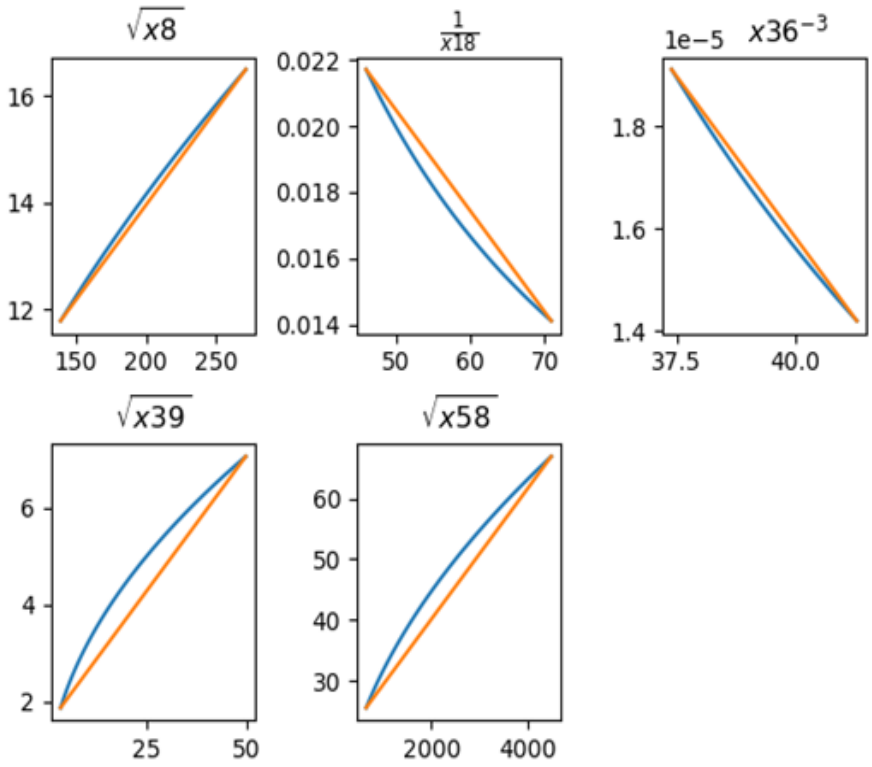

Рис. 3. Визуализация степени нелинейности преобразованных переменных в (8)
Fig. 3. Visualization of the non-linearity degree for the transformed variables in (8)

Заключение

В работе приводится описание ПК КВАРТОН-1, предназначенного для выбора структурных спецификаций квазилинейных регрессий по коэффициенту детерминации и общему критерию нелинейности. С помощью ПК КВАРТОН-1 была построена высококачественная и вполне интерпретируемая квазилинейная регрессионная модель (8) грузовых железнодорожных перевозок в Иркутской области.

Резюмируя, ещё раз подчеркнем, что разработанный нами ПК КВАРТОН-1 является универсальным и может быть использован для решения прикладных задач из самых разных предметный областей.

Список литературы

Баенхаева А.В., Базилевский М.П., Носков С.И. 2016. Выбор структурной спецификации регрессионной модели валового регионального продукта Иркутской области. Информационные технологии и проблемы математического моделирования сложных систем, 16: 31–38.

Базилевский М.П. 2018. Критерии нелинейности квазилинейных регрессионных моделей. Моделирование, оптимизация и информационные технологии, 6 (4): 185–195. DOI: 10.26102/2310-6018/2018.23.4.015

Базилевский М.П., Караулова А.В. 2021. Выбор оптимального соотношения между точностью и нелинейностью при построении квазилинейных регрессионных моделей. Вестник кибернетики, 4 (44): 63–70.

Базилевский М.П., Носков С.И. 2012. Методические и инструментальные средства построения некоторых типов регрессионных моделей. Системы. Методы. Технологии, 1: 80–87.

Носков С.И. 2021. Динамический критерий согласованности поведения при оценке адекватности регрессионных моделей. Вестник Технологического университета, 24 (7): 103–105.

Носков С.И. 2021. Реализация конкурса регрессионных моделей с применением критерия согласованности поведения. Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии, 2: 153–160.

Носков С.И. 1996. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск, Облинформпечать, 321 с.

Носков С.И., Врублевский И.П. 2016. Регрессионная модель динамики эксплуатационных показателей функционирования железнодорожного транспорта. Современные технологии. Системный анализ. Моделирование, 2 (50): 192–197.

Щербанин Ю. А., Ивин Е.А., Курбацкий А.Н., Глазунова А.А. 2017. Эконометрическое моделирование и прогнозирование спроса на грузовые перевозки в России в 1992–2015 гг. Научные труды: Институт народнохозяйственного прогнозирования РАН, 15.

Borowiak D. S. 2020. Model discrimination for nonlinear regression models. CRC Press.

Brook R. J., Arnold G. C. 2018. Applied regression analysis and experimental design. CRC Press.

Cavanaugh J. E., Neath A. A. 2019. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. Wiley Interdisciplinary Reviews: Computational Statistics, 11 (3): e1460. https://doi.org/10.1002/wics.1460

Chicco D., Warrens M. J., Jurman G. 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science 7: e623. https://doi.org/10.7717/peerj-cs.623

Gao H., Zhang M., Goodchild A. 2020. Empirical Analysis of Relieving High-Speed Rail Freight Congestion in China. Sustainability, 12 (23): 9918. https://doi.org/10.3390/su12239918

Khan M. Z., Khan F. N. 2020. Estimating the demand for rail freight transport in Pakistan: A time series analysis. Journal of Rail Transport Planning & Management, 14: 100176. https://doi.org/10.1016/j.jrtpm.2019.100176

Lawrence K. D. 2019. Robust regression: analysis and applications. Routledge.

Lee H. K., Kim H. B. 2018. The impacts of rail freight rate changes on regional economies, modal shift, and environmental quality in Korea. International Journal of Urban Sciences, 22 (4): 517–528. https://doi.org/10.1080/12265934.2018.1476176

Mize T. D. 2019. Best practices for estimating, interpreting, and presenting nonlinear interaction effects. Sociological Science, 6: 81–117. http://dx.doi.org/10.15195/v6.a4

Montgomery D. C., Peck E. A., Vining G. G. 2021. Introduction to linear regression analysis. John Wiley & Sons.

Pompigna A., Mauro R. 2020. Input/Output models for freight transport demand: a macro approach to traffic analysis for a freight corridor. Archives of Transport, 54 (2): 21–42. DOI: 10.5604/01.3001.0014.2729

References

Baenkhaeva A.V., Bazilevskiy M.P., Noskov S.I. 2016. Vybor strukturnoy spetsifikatsii regressionnoy modeli valovogo regional'nogo produkta Irkutskoy oblasti [Choice of the regression model structural specification for the gross regional product of the Irkutsk region]. Informatsionnye tekhnologii i problemy matematicheskogo modelirovaniya slozhnykh sistem, 16: 31–38.

Bazilevskiy M.P. 2018. Kriterii nelineynosti kvazilineynykh regressionnykh modeley [Nonlinear criteria of quasilinear regression models]. Modelirovanie, optimizatsiya i informatsionnye tekhnologii, 6(4): 185–195. DOI: 10.26102/2310-6018/2018.23.4.015

Bazilevskiy M.P., Karaulova A.V. 2021. Vybor optimal'nogo sootnosheniya mezhdu tochnost'yu i nelineynost'yu pri postroenii kvazilineynykh regressionnykh modeley [Selecting the optimum relationship between accuracy and non-linearity in constructing quasi-linear regression models]. Vestnik kibernetiki, 4 (44): 63–70.

Bazilevskiy M.P., Noskov S.I. 2012. Metodicheskie i instrumental'nye sredstva postroeniya nekotorykh tipov regressionnykh modeley [Methodology and instrumental tools for construction some types of regression models]. Sistemy. Metody. Tekhnologii, 1: 80–87.

Noskov S.I. 2021. Dinamicheskiy kriteriy soglasovannosti povedeniya pri otsenke adekvatnosti regressionnykh modeley [Dynamic criterion for consistency of behavior when evaluating the adequacy of regression models]. Vestnik Tekhnologicheskogo universiteta, 24 (7): 103–105.

Noskov S.I. 2021. Realizatsiya konkursa regressionnykh modeley s primeneniem kriteriya soglasovannosti povedeniya [Implementation of the competition of regression models using the criterion of consistency

of behavior]. Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Sistemnyy analiz i informatsionnye tekhnologii, 2: 153–160.

Noskov S.I. 1996. Tekhnologiya modelirovaniya ob"ektov s nestabil'nym funktsionirovaniem i neopredelennost'yu v dannykh [Object modeling technology with unstable operation and data uncertainty]. Irkutsk, Oblinformpechat', 321 p.

Noskov S.I., Vrublevskiy I.P. 2016. Regressionnaya model' dinamiki ekspluatatsionnykh pokazateley funktsionirovaniya zheleznodorozhnogo transporta [Railway transport functioning the regression model performance indicators dynamics]. Sovremennye tekhnologii. Sistemnyy analiz. Modelirovanie, 2 (50): 192–197.

Shcherbanin Yu. A., Ivin E.A., Kurbatskiy A.N., Glazunova A.A. 2017. Ekonometricheskoe modelirovanie i prognozirovanie sprosa na gruzovye perevozki v Rossii v 1992–2015 gg [Econometric Modeling and Forecasting of Demand for Freight Transportation in Russia in 1992–2015]. Nauchnye trudy: Institut narodnokhozyaystvennogo prognozirovaniya RAN, 15.

Borowiak D. S. 2020. Model discrimination for nonlinear regression models. CRC Press.

Brook R. J., Arnold G. C. 2018. Applied regression analysis and experimental design. CRC Press.

Cavanaugh J. E., Neath A. A. 2019. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. Wiley Interdisciplinary Reviews: Computational Statistics, 11 (3): e1460. https://doi.org/10.1002/wics.1460

Chicco D., Warrens M. J., Jurman G. 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science 7: e623. https://doi.org/10.7717/peerj-cs.623

Gao H., Zhang M., Goodchild A. 2020. Empirical Analysis of Relieving High-Speed Rail Freight Congestion in China. Sustainability, 12 (23): 9918. https://doi.org/10.3390/su12239918

Khan M. Z., Khan F. N. 2020. Estimating the demand for rail freight transport in Pakistan: A time series analysis. Journal of Rail Transport Planning & Management, 14: 100176. https://doi.org/10.1016/j.jrtpm.2019.100176

Lawrence K. D. 2019. Robust regression: analysis and applications. Routledge.

Lee H. K., Kim H. B. 2018. The impacts of rail freight rate changes on regional economies, modal shift, and environmental quality in Korea. International Journal of Urban Sciences, 22 (4): 517–528. https://doi.org/10.1080/12265934.2018.1476176

Mize T. D. 2019. Best practices for estimating, interpreting, and presenting nonlinear interaction effects. Sociological Science, 6: 81–117. http://dx.doi.org/10.15195/v6.a4

Montgomery D. C., Peck E. A., Vining G. G. 2021. Introduction to linear regression analysis. John Wiley & Sons.

Pompigna A., Mauro R. 2020. Input/Output models for freight transport demand: a macro approach to traffic analysis for a freight corridor. Archives of Transport, 54 (2): 21–42. DOI: 10.5604/01.3001.0014.2729

Конфликт интересов: о потенциальном конфликте интересов не сообщалось.
Conflict of interest: no potential conflict of interest related to this article was reported.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Караулова Анна Витальевна, аспирант кафедры информационных систем и защиты информации, Иркутский государственный университет путей сообщения, г. Иркутск, Россия

Базилевский Михаил Павлович, кандидат технических наук, доцент кафедры математики, Иркутский государственный университет путей сообщения, г. Иркутск, Россия

INFORMATION ABOUT THE AUTHORS

Anna V. Karaulova, Postgraduate Student of the Department of Information Systems and Information Protection, Irkutsk State Transport University, Irkutsk, Russia

Mikhail P. Bazilevskiy, Candidate of Technical Sciences, Associate Professor of the Department of Mathematics, Irkutsk State Transport University, Irkutsk, Russia