УДК 004.622+004.891.3

О ВОЗМОЖНОСТИ ПОВЫШЕНИЯ СТАТИСТИЧЕСКОЙ ЗНАЧИМОСТИ РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ В ЗАДАЧАХ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ВЫБОРОК МАЛОГО ОБЪЕМА

ГРАЧЕВ Владимир Васильевич, докт. техн. наук, профессор[1]; e-mail: v_grach@mail.ru
ШВАРЦ Михаил Александрович, канд. техн. наук, доцент[2]; e-mail: shvarts4545@mai.ru
ГРИЩЕНКО Александр Васильевич, докт. техн. наук, профессор[1]; e-mail: sanklok@mail.ru
ШВАРЦ Филипп Михайлович, инженер, магистр[3]; e-mail: films@mail.ru

[1] Петербургский государственный университет путей сообщения Императора Александра I, кафедра «Локомотивы и локомотивное хозяйство», Санкт-Петербург
[2] Петербургский государственный университет путей сообщения Императора Александра I, кафедра «Высшая математика», Санкт-Петербург
[3] Петербургский государственный университет путей сообщения Императора Александра I, кафедра «Химия», Санкт-Петербург

Рассмотрена проблема повышения достоверности и статистической значимости регрессионных моделей объектов исследований, построенных на экспериментальных выборках данных небольшого объема. Это вынуждает исследователя использовать линейные модели с минимальным количеством варьируемых факторов, однако даже при таком выборе вида модели недостаточная статистическая значимость оценок параметров исключает возможность использования ее для достоверного прогнозирования изменения объясняемых переменных. С целью расширения возможности выбора вида модели на стадии спецификации и повышения статистической значимости оценок ее параметров предлагается расширить объем экспериментальных данных с помощью статистической модели объекта исследования, построенной на основе генеративно-состязательной нейронной сети. При обучении на выборке небольшого объема, полученной в ходе экспериментального исследования объекта, генератор условной генеративно-состязательной сети генерирует кластеры данных с центроидами, соответствующими точкам обучающей (экспериментальной) выборки. Приведены результаты анализа данных физического эксперимента, подтверждающие основные ее положения.

Ключевые слова: экспериментальная выборка, линейная модель регрессии, статистическая значимость, оценка параметров регрессии, условная генеративно-состязательная сеть, статистическая модель, множественная корреляция, расстояние Евклида – Махалонобиса.

DOI: 10.20295/2412-9186-2024-10-04-382-394

▼ Введение

Решение целого ряда задач, связанных с анализом экспериментальных данных (изучение свойств материалов, подбор компонентов смесей, синтез новых веществ) или данных мониторинга (анализ надежности и диагностирование сложных технических систем), предполагает построение регрессионной эталонной модели исследуемого процесса. Одним из основных факторов, определяющих достоверность модели и показатели ее качества, является объем выборки данных, на которой строится модель. Очень часто этот объем ограничен временными и материальными ресурсами исследователя или режимами эксплуатации исследуемой системы, что вынуждает использовать для описания исследуемого процесса простые линейные модели с ограниченным количеством варьируемых параметров.

Возможным решением проблемы представляется генерация дополнительного объема данных с помощью статистической модели процесса, построенной с использованием генеративной состязательной нейронной сети GAN (Generative Adversarial Network) [1].

Ранее авторами решалась аналогичная задача для моделей машинного обучения [2] с ограниченным объемом обучающей выборки,

недостаточным для качественного обучения многомерного интеллектуального классификатора, однако позволяющим сформировать статистическую модель исследуемого объекта на основе сети GAN.

Особенностью задачи, которой посвящена настоящая статья, является сверхмалый объем исходной выборки, недостаточный для построения даже статистически значимой линейной регрессионной модели.

1. Постановка задачи

Практически любой объект исследования (ОИ) может быть представлен отображением вида:

$$H : \overline{X}(t) \xrightarrow{\overline{R}(t)} \overline{Y}(t), \qquad (1)$$

где $\overline{X} = \{x_1, x_2, x_3, ..., x_m\}$ — вектор независимых входных параметров ОИ, определяемый порядком его взаимодействия с окружающей средой;

$\overline{Y} = \{y_1, y_2, y_3, ..., y_l\}$ — вектор зависимых выходных параметров ОИ, характеризующих его реакцию на входное воздействие $\overline{X}$;

$\overline{R} = \{r_1, r_2, r_3, ..., r_n\}$ — вектор структурных параметров ОИ, характеризующий сущность образующих его физических явлений или процессов.

Измерительная информация, регистрируемая в ходе исследовательского эксперимента или мониторинга процесса, может быть представлена векторами $\overline{Z_x} = \{z_{x1}, z_{x2}, \ldots, z_{xr}\}$ и $\overline{Z_y} = \{z_{y1}, z_{y2}, \ldots, z_{yp}\}$ измеренных значений входных и выходных параметров, при этом, как правило, $r \ll m$ и $p \ll l$.

В результате эксперимента образуется совокупность пар векторов измеренных значений входных и выходных параметров:

$$\overline{Z} = \left\{(\overline{Z_x^1}, \overline{Z_y^1}), (\overline{Z_x^2}, \overline{Z_y^2}), (\overline{Z_x^3}, \overline{Z_y^3}), \ldots, (\overline{Z_x^k}, \overline{Z_y^k})\right\}. \qquad (2)$$

После нормализации значений компонент векторов и приведения их к интервалу [0, 1] эта совокупность может рассматриваться как $(r+p)$-мерное распределение k точек в области пространства нормированных контролируемых параметров ОИ. Закон распределения $P(z_{x1}, z_{x2}, ..., z_{xr}, z_{y1}, z_{y2}, ..., z_{yp})$ определяется как физическими основами протекания изучаемого процесса (параметрами отображения H и компонентов вектора $\overline{R}$ (1)), так и способом измерения компонентов векторов $\overline{X}$ и $\overline{Y}$.

При достаточном объеме выборки $\overline{Z}$ генератор (generator) сети GAN, обученной на этой выборке, генерирует данные в соответствии с многомерным распределением $P(z_{x1}, z_{x2}, ..., z_{xr}, z_{y1}, z_{y2}, ..., z_{yp})$, заданным выборкой. Необходимость аугментации опытных данных в таких задачах обусловлена, как правило, выбором интеллектуальных эталонных моделей процессов, для обучения которых необходимы выборки больших объемов [2].

Однако часто при решении задач небольшой размерности располагаемый объем экспериментальных данных недостаточен не только для определения параметров их распределения, но и для построения статистически значимых регрессионных зависимостей.

Целью работы является исследование возможности использования сетей GAN (Generative Adversarial Network) для повышения статистической значимости регрессионных зависимостей в задачах обработки экспериментальных выборок небольшого объема.

2. Выбор метода решения

Анализ различных вариантов конфигурации сетей GAN показал, что для решения задачи может быть применена условная (conditional) сеть GAN (CGAN) [3], в которой оценка соответствия распределения обучающей выборки и сгенерированных данных осуществляется с использованием метрики Вассерштейна (Wassershtein) [4] в сочетании с параметрическим (посредством штрафной функции (Gradient Penalty) ограничением величины градиента критика (CWGAN-GP) [5].

При обучении сети CWGAN-GP на выборке небольшого объема каждому вектору обучающей выборки ставится в соответствие вектор категорированных признаков, определяющий точку в пространстве бинарных признаков, размерность которого равна объему выборки. Этот вектор добавляется как к случайному вектору, подаваемому на вход генератора, так и к векторам обучающей выборки и сгенерированных данных, подаваемых на вход критика. После завершения обучения сети обученный генератор генерирует данные в соответствии с заданным вектором признаков,

определяющим центр (математическое ожидание) кластера данных [2].

Таким образом, результатом генерации являются k кластеров точек в пространстве измеренных параметров процесса, каждый из которых соответствует одному из векторов обучающей выборки $(\overline{Z_x^k}, \overline{Z_y^k})$ (рис. 1).

Необходимо подчеркнуть, что многомерные распределения $(p'_{zx1}, p'_{zy1}), (p'_{zx2}, p'_{zy2}), ..., (p'_{zxk}, p'_{zyk})$ не являются независимыми. Между сгенерированными векторами $(\overline{Z_x^{1'}}, \overline{Z_x^{2'}}, ..., \overline{Z_x^{k'}})$ входных и $(\overline{Z_y^{1'}}, \overline{Z_y^{2'}}, ..., \overline{Z_y^{k'}})$ выходных параметров ОИ существует множественная корреляционная связь, то есть они могут рассматриваться в качестве расширенной статистической модели физического эксперимента. Качество модели может оцениваться расстоянием между центрами кластеров $(\overline{Z_x^{k'}}, \overline{Z_y^{k'}})$ сгенерированных данных и соответствующими им векторами $(\overline{Z_x^k}, \overline{Z_y^{k'}})$ измеренных значений параметров ОИ.

Эта модель не дает дополнительных (по отношению к реальному эксперименту) знаний о сущности ОИ, однако позволяет повысить качество и достоверность результатов обработки экспериментальных данных.

В частности, увеличение объема корреляционного поля существенно расширяет возможности выбора структуры регрессионной модели ОИ на стадии ее спецификации.

3. Обсуждение результатов применения метода

В табл. 1 приведены результаты сложного и ресурсоемкого физического эксперимента.

Здесь x_1, x_2, x_3, x_4 — независимые параметры (регрессоры), характеризующие внешнее воздействие на ОИ, y_1, y_2, y_3, y_4 — объясняющие параметры (регрессанты), характеризующие реакцию ОИ на внешнее воздействие.

Задача исследования, в рамках которого проводился эксперимент, состоит в определении вектора $\overline{X}^0 = \{x_1^0, x_2^0, x_3^0, x_4^0\}$, доставляющего минимум некоторому критерию $\overline{\Phi}^0 = \{y_1^0, y_2^0, y_3^0, y_4^0\}$.

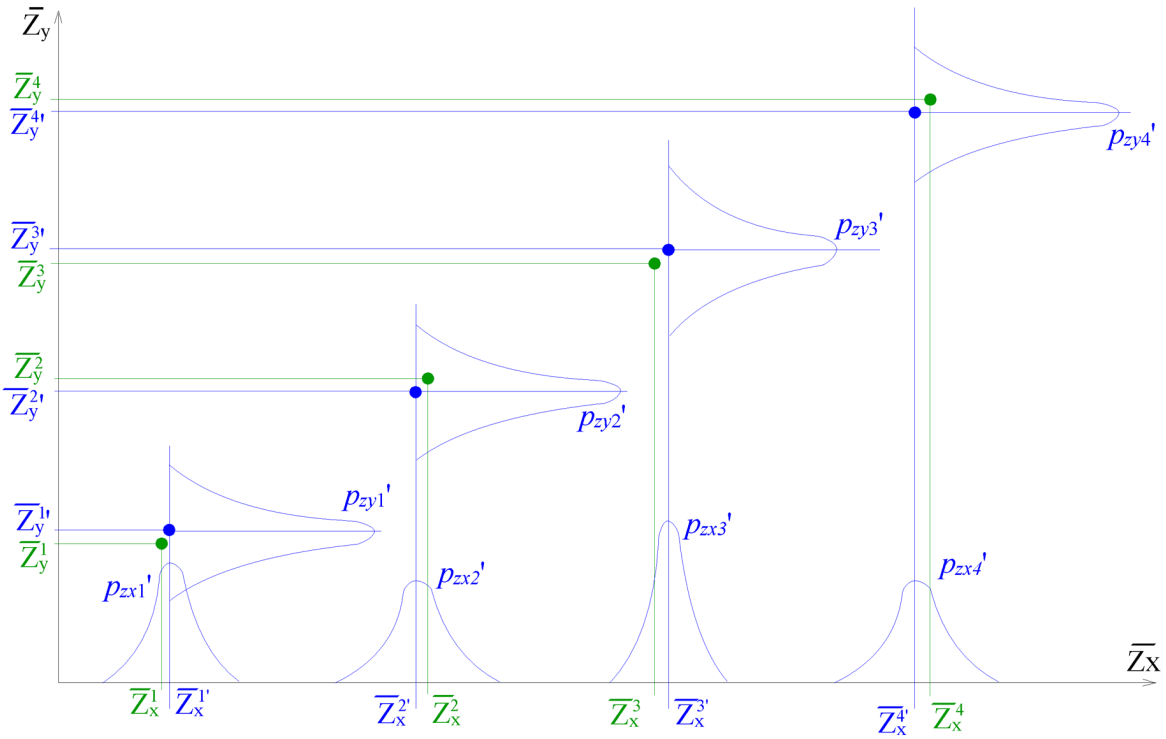

Рис. 1. Результаты эксперимента и его статистическая модель: $(\overline{Z_x^1}, \overline{Z_y^1}), ..., (\overline{Z_x^k}, \overline{Z_y^k})$ — векторы результатов измерений параметров процесса; $(\overline{Z_x^{1'}}, \overline{Z_y^{1'}}), ..., (\overline{Z_x^{k'}}, \overline{Z_y^{k'}})$ — векторы математического ожидания модельных (сгенерированных) значений параметров процесса; $(p'_{zx1}, p'_{zy1}), (p'_{zx2}, p'_{zy2}), ..., (p'_{zxk}, p'_{zyk})$ — распределения модельных (сгенерированных) значений параметров процесса

Таблица 1. Экспериментальные данные

№ изм.	x_1	x_2	x_3	x_4	y_1	y_2	y_3	y_4
1	0	0	0	0	360	25,0	42,00	4,9
2	0,5	0	0	0	564	18,0	53,50	6,2
3	0,6	0	0	0	568	17,0	54,30	6,4
4	0,7	0	0	0	572	17,0	54,00	6,4
5	0,6	0,1	0	0	590	16,0	55,80	6,6
6	0,6	0,2	0	0	595	15,5	56,70	6,8
7	0,6	0,3	0	0	600	15,0	56,70	6,8
8	0,6	0,2	0,5	0	615	14,0	61,10	7,3
9	0,6	0,2	0,6	0	622	13,5	62,00	7,6
10	0,6	0,2	0,7	0	624	13,5	62,00	7,6
11	0,6	0,2	0,6	0,3	633	12,0	68,10	8,3
12	0,6	0,2	0,6	0,5	638	11,5	69,00	8,5
13	0,6	0,2	0,6	0,7	634	11,5	69,00	8,5

Для организации поиска в пространстве признаков $\{x_1, x_2, x_3, x_4\}$ требуется построить регрессионные зависимости вида:

$$\begin{aligned} y_1 &= f_1(x_1, x_2, x_3, x_4), \\ y_2 &= f_2(x_1, x_2, x_3, x_4), \\ y_3 &= f_3(x_1, x_2, x_3, x_4), \\ y_4 &= f_4(x_1, x_2, x_3, x_4). \end{aligned} \quad (3)$$

Ограниченный объем экспериментальных данных вынуждает принять линейный тип модели с минимальным количеством влияющих факторов вида:

$$y = A + B \cdot x_1 + C \cdot x_2 + D \cdot x_3 + E \cdot x_4. \quad (4)$$

Обычно рекомендуется придерживаться правила, согласно которому число факторов в 4–6 раз меньше объема статистического материала [6–9]. При нарушении этого условия число степеней свободы остаточной дисперсии существенно уменьшается, это приводит к тому, что оценки значений коэффициентов в уравнениях регрессии становятся статистически незначимыми.

Результаты определения оценок коэффициентов моделей с использованием метода наименьших квадратов [10], а также показатели качества регрессионных зависимостей приведены в табл. 2–5.

Анализ полученных результатов показывает, что ряд оценок параметров уравнений регрессии не являются статистически значимыми для уровня значимости 5 % (C и E в уравнении для y_1 (табл. 1), C в уравнении для y_3 (табл. 4)), что не позволяет осуществлять достоверное прогнозирование изменения зависимых переменных при изменении значений регрессоров.

Для повышения статистической значимости регрессионных моделей в целом и оценок их параметров необходимо увеличить объем экспериментального материала.

Для решения задачи была сформирована сеть CWGAN-GP [5] со следующей структурой (рис. 2):

- генератор (8+8)–380–520–400–370–340–240–8 с функцией активации нейронов скрытых слоев *‘selu’*, выходного слоя — *‘sigmoid’*;
- критик (8+8)–333–366–455–333–233–133–1 с функцией активации нейронов скрытых слоев *‘selu’*, выходного слоя — *‘sigmoid’*.

Генератор формирует случайные векторы Z из заданного распределения $p(Z)$ (как правило, нормального, вида $N(0, 1)$) и генерирует из них объекты $X_p = G(Z)$, которые подаются на вход второй сети (критик). Вместе с ними на вход критика подаются объекты X_S из имеющейся выборки. Выходом критика является

Таблица 2. Значения параметров линейного уравнения регрессии для y_1

Параметры	Коэффициенты	Стандартная ошибка	*t*-статистика ($t_{кр}^{\alpha=0,05} = 2,306$)	*P*-значение
A	369,648	15,278	24,195	9,080E-09
B	329,446	28,707	11,476	3,010E-06
C	125,947	55,105	2,2856	0,0516
D	48,291	19,856	2,432	0,0411
E	24,676	23,175	1,065	0,318

Таблица 3. Значения параметров линейного уравнения регрессии для y_2

Параметры	Коэффициенты	Стандартная ошибка	*t*-статистика ($t_{кр}^{\alpha=0,05} = 2,306$)	*P*-значение
A	24,753	0,511	48,406	3,670E-11
B	-12,421	0,961	-12,927	1,210E-06
C	-8,472	1,844	-4,594	0,00177
D	-3,417	0,665	-5,142	0,000883
E	-3,533	0,776	-4,555	0,00186

Таблица 4. Значения параметров линейного уравнения регрессии для y_3

Параметры	Коэффициенты	Стандартная ошибка	*t*-статистика ($t_{кр}^{\alpha=0,05} = 2,306$)	*P*-значение
A	42,551	1,424	29,903	1,700E-09
B	19,041	2,674	7,121	9,990E-05
C	11,288	5,132	2,199	0,0590
D	9,881	1,849	5,343	0,000692
E	12,033	2,158	5,575	0,000526

Таблица 5. Значения параметров линейного уравнения регрессии для y_4

Параметры	Коэффициенты	Стандартная ошибка	*t*-статистика ($t_{кр}^{\alpha=0,05} = 2,306$)	*P*-значение
A	4,944	0,158	31,233	1,200E-09
B	2,335	0,297	7,849	5,010E-05
C	1,802	0,571	3,157	0,0135
D	1,425	0,206	6,927	0,000121
E	1,625	0,240	6,767	0,000143

вероятность $D(X)$ принадлежности входного объекта реальной выборке.

Генератор и критик обучаются отдельно, но в рамках одной сети. После завершения обучения генератор может использоваться для генерации искусственных данных с распределением, соответствующим исходной выборке, из векторов с нормально распределенными значениями компонентов, подаваемых на его вход.

Для каждого из векторов экспериментальной выборки (табл. 1) был сгенерирован кластер модельных данных объемом 200 векторов.

С целью подтверждения наличия в модельных данных функциональной зависимости между входными (x_1, x_2, x_3, x_4) и выходными

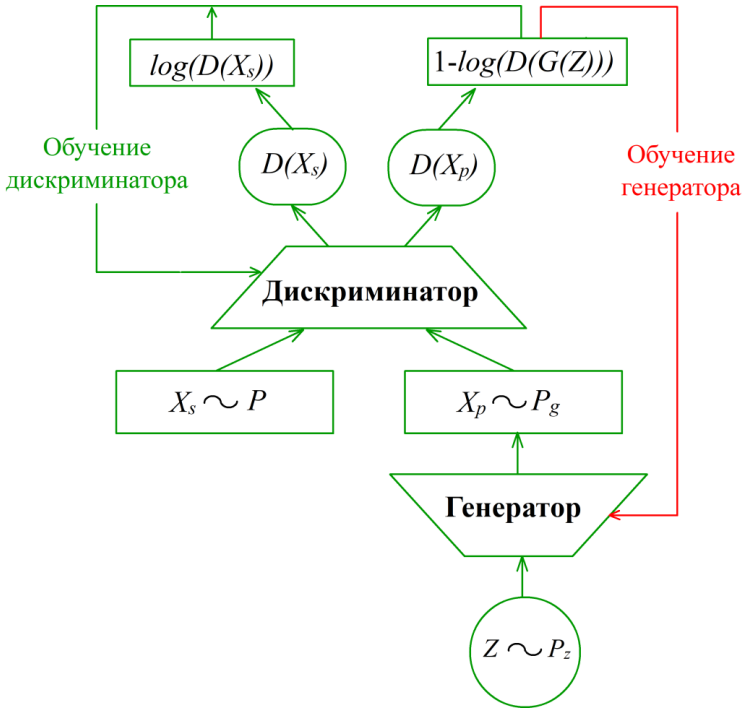

Рис. 2. Схема структуры сети *CWGAN–GP*

(y_1, y_2, y_3, y_4) признаками, обусловленной свойствами ОИ, для каждого из кластеров модельных данных были определены множественные коэффициенты корреляции R_{X,y_n}, $n = 1...4$ [11] между вектором входных признаков $\{x_1, x_2, x_3, x_4\}$ и каждым из выходных признаков (y_1, y_2, y_3, y_4).

В качестве базы для сравнения использовались массивы случайных чисел с математическими ожиданиями, совпадающими с измеренными значениями входных и выходных параметров ОИ (табл. 1) (104 массива по 200 элементов в каждом), сгенерированные с использованием соответствующих функций пакета MS Excel [12]. Объединение массивов, соответствующих одной строчке табл. 1, составляет кластер случайных данных, который может быть поставлен в соответствие кластеру модельных данных, сгенерированных сетью CWGAN-GP.

Результаты определения множественных коэффициентов корреляции между входными признаками ОИ и каждым из его выходных признаков для кластеров модельных и случайных данных приведены в табл. 6.

Как следует из табл. 6, несмотря на невысокий уровень множественной корреляции между входными и выходными данными внутри модельных кластеров, обусловленный наличием случайной составляющей в данных, он кратно превосходит уровень множественной корреляции между соответствующими параметрами в кластерах случайных данных.

В качестве одного из показателей адекватности такой модели может рассматриваться расстояние между центроидами сгенерированных кластеров и соответствующими им точками экспериментальных данных в пространстве признаков. Используя обобщенную метрику Евклида – Махалонобиса [13], можно определить это расстояние по формуле:

$$d = \sqrt{\sum (X - \mu)(S + E)^{-1}(X - \mu)^T}, \qquad (5)$$

где S — ковариационная матрица нормализованных модельных векторов кластера;

E — единичная матрица;

μ — нормализованный вектор экспериментальных данных, соответствующий данному кластеру;

X — центроид нормализованных модельных векторов кластера.

В качестве примера рассмотрим последнюю строку табл. 6. Значения компонент векторов X и μ для этой строки приведены в табл. 7.

Ковариационная матрица S 200 модельных векторов, соответствующих этой точке экспериментальных данных, приведена в табл. 8.

С учетом приведенных данных расстояние между экспериментальной точкой, соответствующей строке 13 табл. 1, и центроидом соответствующего ей кластера сгенерированных данных в пространстве нормализованных значений входных и выходных признаков ОИ составляет $d_{13} = 0{,}0430$.

Соответствующие расстояния для точек 1–12 табл. 1 приведены в табл. 9.

Таблица 6. Множественные коэффициенты корреляции между векторами входных и выходными параметрами в модельных и случайных кластерах данных

Кластер	y_1		y_2		y_3		y_4	
	CWGAN-GP	*Random*	*CWGAN-GP*	*Random*	*CWGAN-GP*	*Random*	*CWGAN-GP*	*Random*
0	0,416	0,093	0,378	0,073	0,398	0,101	0,252	0,081
1	0,240	0,116	0,220	0,090	0,310	0,052	0,330	0,123
2	0,479	0,112	0,298	0,097	0,178	0,087	0,173	0,108
3	0,258	0,090	0,478	0,077	0,184	0,106	0,295	0,109
4	0,438	0,095	0,361	0,118	0,182	0,094	0,254	0,056
5	0,431	0,100	0,311	0,051	0,328	0,116	0,264	0,071
6	0,434	0,093	0,254	0,084	0,384	0,06	0,362	0,107
7	0,337	0,073	0,285	0,102	0,243	0,091	0,289	0,085
8	0,36	0,077	0,276	0,094	0,234	0,114	0,181	0,108
9	0,464	0,108	0,387	0,063	0,402	0,105	0,212	0,082
10	0,237	0,058	0,286	0,108	0,302	0,118	0,206	0,075
11	0,237	0,098	0,329	0,042	0,42	0,069	0,296	0,089
12	0,185	0,111	0,264	0,115	0,235	0,087	0,202	0,072

Таблица 7. Нормализованные значения компонент вектора экспериментальных данных и вектора центроида кластера

	x_1	x_2	x_3	x_4	y_1	y_2	y_3	y_4
X	0,8502	0,6313	0,8358	1,0000	1,0000	0,4576	1,0000	1,0000
μ	0,8571	0,6666	0,8571	1,0000	0,9930	0,4600	1,0000	1,0000

Таблица 8. Ковариационная матрица *S* векторов модельных данных

	x_1	x_2	x_3	x_4	y_1	y_2	y_3	y_4
x_1	0,000345	0,000124	0,000109	$1{,}145 \cdot 10^{-05}$	$-8{,}120 \cdot 10^{-06}$	$1{,}522 \cdot 10^{-05}$	$-2{,}257 \cdot 10^{-06}$	$-4{,}991 \cdot 10^{-06}$
x_2	0,000124	0,00215	0,000451	0,000275	$1{,}637 \cdot 10^{-05}$	0,000276	$3{,}828 \cdot 10^{-05}$	$7{,}742 \cdot 10^{-06}$
x_3	0,000109	0,000451	0,000888	0,000872	$3{,}504 \cdot 10^{-06}$	$-4{,}497 \cdot 10^{-05}$	$4{,}569 \cdot 10^{-05}$	$1{,}216 \cdot 10^{-05}$
x_4	$1{,}145 \cdot 10^{-05}$	0,000275	0,000872	0,00152	$-1{,}196 \cdot 10^{-06}$	$9{,}117 \cdot 10^{-05}$	$1{,}598 \cdot 10^{-05}$	$-5{,}850 \cdot 10^{-06}$
y_1	$-8{,}120 \cdot 10^{-06}$	$1{,}637 \cdot 10^{-05}$	$3{,}504 \cdot 10^{-06}$	$-1{,}196 \cdot 10^{-06}$	$1{,}015 \cdot 10^{-06}$	$4{,}215 \cdot 10^{-06}$	$2{,}390 \cdot 10^{-06}$	$9{,}072 \cdot 10^{-07}$
y_2	$1{,}522 \cdot 10^{-05}$	0,000276	$-4{,}497 \cdot 10^{-05}$	$-9{,}117 \cdot 10^{-05}$	$4{,}215 \cdot 10^{-06}$	$7{,}506 \cdot 10^{-05}$	$4{,}575 \cdot 10^{-06}$	$-8{,}525 \cdot 10^{-07}$
y_3	$-2{,}257 \cdot 10^{-06}$	$3{,}828 \cdot 10^{-05}$	$4{,}569 \cdot 10^{-05}$	$1{,}598 \cdot 10^{-05}$	$2{,}390 \cdot 10^{-06}$	$4{,}575 \cdot 10^{-06}$	$8{,}798 \cdot 10^{-06}$	$3{,}474 \cdot 10^{-06}$
y_4	$-4{,}991 \cdot 10^{-06}$	$7{,}742 \cdot 10^{-06}$	$1{,}216 \cdot 10^{-05}$	$-5{,}850 \cdot 10^{-06}$	$9{,}072 \cdot 10^{-07}$	$-8{,}525 \cdot 10^{-07}$	$3{,}474 \cdot 10^{-06}$	$1{,}944 \cdot 10^{-06}$

Таблица 9. Результаты вычислений расстояний между экспериментальными точками и цетроидами кластеров модельных данных

d_1	d_2	d_3	d_4	d_5	d_6
0,03489	0,02964	0,02057	0,02979	0,02199	0,03614
d_7	d_8	d_9	d_{10}	d_{11}	d_{12}
0,03632	0,04365	0,02054	0,03452	0,03329	0,02350

Как следует из табл. 9, смещение центров кластеров сгенерированных модельных данных относительно соответствующих им экспериментальных точек в пространстве нормализованных (приведенных к единице) признаков, не превышает 0,044, что наряду с наличием корреляции между векторами входных и выходных признаков свидетельствует об адекватности статистической модели ОИ.

Многократное расширение корреляционного поля существенно расширяет возможности выбора типа регрессионных моделей с учетом сложности описываемого явления. В данном случае для уточнения аппроксимации результатов эксперимента была выбрана полиномиальная модель вида:

$$y = A \cdot x_4^2 + B \cdot x_2^2 + C \cdot x_3^2 + D \cdot x_1^2 + \\ + E \cdot x_1 + G \cdot x_2 + H \cdot x_3 + F \cdot x_4 + K. \quad (6)$$

Значения параметров моделей определялись методом наименьших квадратов с использованием всего массива сгенерированных данных (2600 элементов). Результаты определения оценок параметров приведены в табл. 10–13.

Таблица 10. Значения оценок параметров полиномиального уравнения регрессии для y_1

Параметры	Значение	Стандартная ошибка	*t*-статистика ($t_{кр}^{\alpha=0,01} = 2,578$)	*P*-значение
K	350,716	2,132	164,509	$5,200 \cdot 10^{-147}$
E	745,420	17,977	41,465	$1,560 \cdot 10^{-74}$
G	105,662	7,604	13,895	$4,790 \cdot 10^{-27}$
H	28,867	2,692	10,722	$2,210 \cdot 10^{-19}$
F	44,569	3,0278	14,720	$5,460 \cdot 10^{-29}$
D	–617,966	26,927	–22,950	$1,830 \cdot 10^{-46}$

Таблица 11. Значения оценок параметров полиномиального уравнения регрессии для y_2

Параметры	Значение	Стандартная ошибка	*t*-статистика ($t_{кр}^{\alpha=0,01} = 2,578$)	*P*-значение
K	25,00617	0,0549	455,598	$5,800 \cdot 10^{-199}$
E	–21,677	0,4678	–46,333	$2,720 \cdot 10^{-79}$
G	–13,937	0,725	–19,235	$9,490 \cdot 10^{-39}$
H	–2,58049	0,080081	–32,2233	$1,620 \cdot 10^{-61}$
F	–8,08942	0,289661	–27,9272	$7,790 \cdot 10^{-55}$
D	14,92072	0,698447	21,3627	$4,810 \cdot 10^{-43}$
B	23,07348	2,560997	9,009572	$3,470 \cdot 10^{-15}$
A	6,236561	0,438175	14,23303	$1,090 \cdot 10^{-27}$

Таблица 12. Значения оценок параметров полиномиального уравнения регрессии для y_3

Параметры	Значение	Стандартная ошибка	*t*-статистика ($t_{кр}^{\alpha=0,01} = 2,578$)	*P*-значение
K	42,55811	0,169893	250,5002	$2,600 \cdot 10^{-167}$
E	40,24013	1,43782	27,9869	$6,210 \cdot 10^{-55}$
G	10,38297	0,611184	16,98827	$6,060 \cdot 10^{-34}$
H	15,45864	1,2709	12,16354	$8,810 \cdot 10^{-23}$
F	29,48171	0,896543	32,88375	$1,760 \cdot 10^{-62}$
D	–33,2197	2,154636	–15,4178	$2,000 \cdot 10^{-30}$
C	–12,1406	1,985785	–6,11376	$1,210 \cdot 10^{-08}$
A	–27,3087	1,354339	–20,1639	$1,180 \cdot 10^{-40}$

Таблица 13. Значения оценок параметров полиномиального уравнения регрессии для y_4

Параметры	Значение	Стандартная ошибка	*t*-статистика ($t_{кр}^{\alpha=0,01} = 2,578$)	*P*-значение
K	4,922512	0,022737	216,4965	$1,100 \cdot 10^{-160}$
E	4,455264	0,19243	23,15261	$1,160 \cdot 10^{-46}$
G	1,705444	0,081134	21,02016	$1,560 \cdot 10^{-42}$
H	1,130765	0,029399	38,46225	$2,040 \cdot 10^{-70}$
F	3,754787	0,119973	31,29705	$2,010 \cdot 10^{-60}$
D	–3,23097	0,288348	–11,2051	$1,630 \cdot 10^{-20}$
A	–3,30645	0,181339	–18,2336	$9,070 \cdot 10^{-37}$

Коэффициенты, не указанные в табл. 10–13, в соответствующих уравнениях регрессии равны нулю.

Как следует из табл. 10–13, оценки всех параметров полиномиальных уравнений множественной регрессии выходных переменных y_1, y_2, y_3, y_4 значимы, причем не только на уровне значимости 0,05, но и на уровне значимости 0,01. Коэффициент детерминации R для каждого уравнения регрессии выше 0,99, что свидетельствует о высокой адекватности моделей, построенных на сгенерированных данных.

Этот вывод подтверждается графическими представлениями модельных и экспериментальных значений выходных параметров ОИ, приведенными на рис. 3–6.

Выводы

Полученные результаты позволяют сделать следующие выводы:

1. Генератор сети GAN, обученный на экспериментальной выборке многомерных данных, полученной в результате исследования физического объекта или явления, может рассматриваться в качестве статистической модели объекта исследования.

2. При ограниченном объеме экспериментальной выборки для построения генерирующей модели целесообразно использовать сети CGAN.

3. Несмотря на наличие случайной составляющей в данных, сгенерированных обученным генератором сети CGAN, между векторами входных и выходных признаков этих данных присутствует устойчивая множественная корреляционная связь, обусловленная физическими принципами функционирования объекта и связью между векторами входных и выходных признаков экспериментальной выборки. В массиве данных, полученном случайной генерацией с центрами в точках, соответствующих векторам экспериментальной выборки, множественная корреляция между векторами входных и выходных признаков отсутствует.

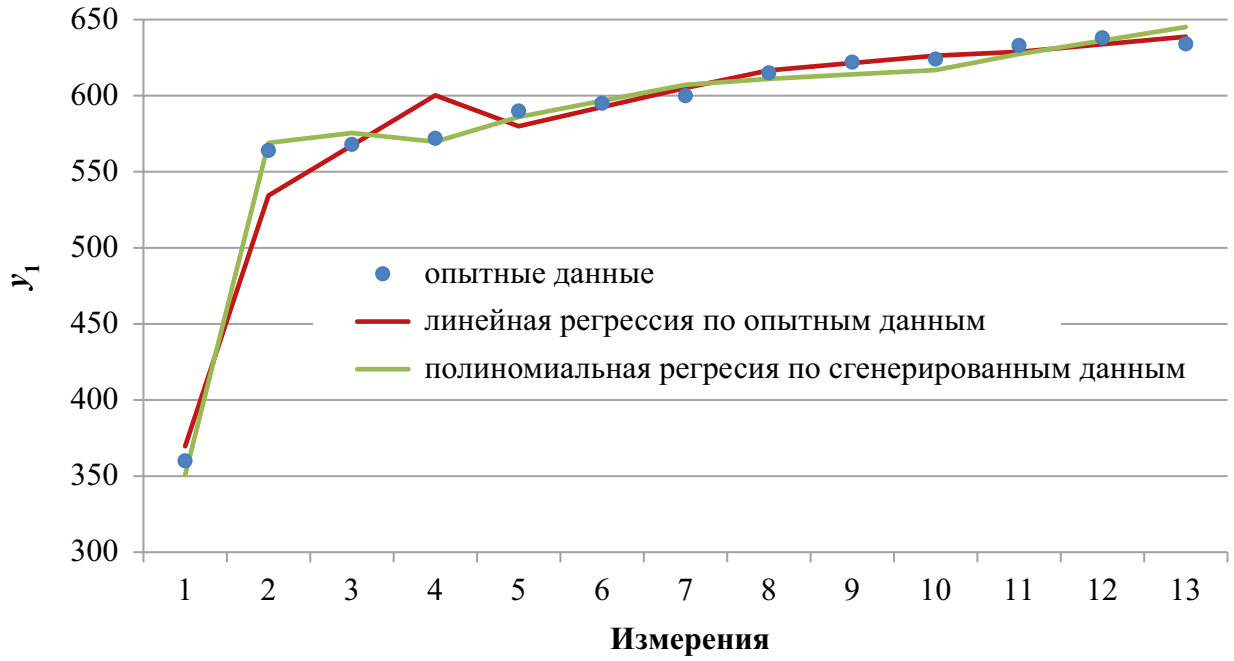

Рис. 3. Графическое представление экспериментальных и модельных значений параметра y_1

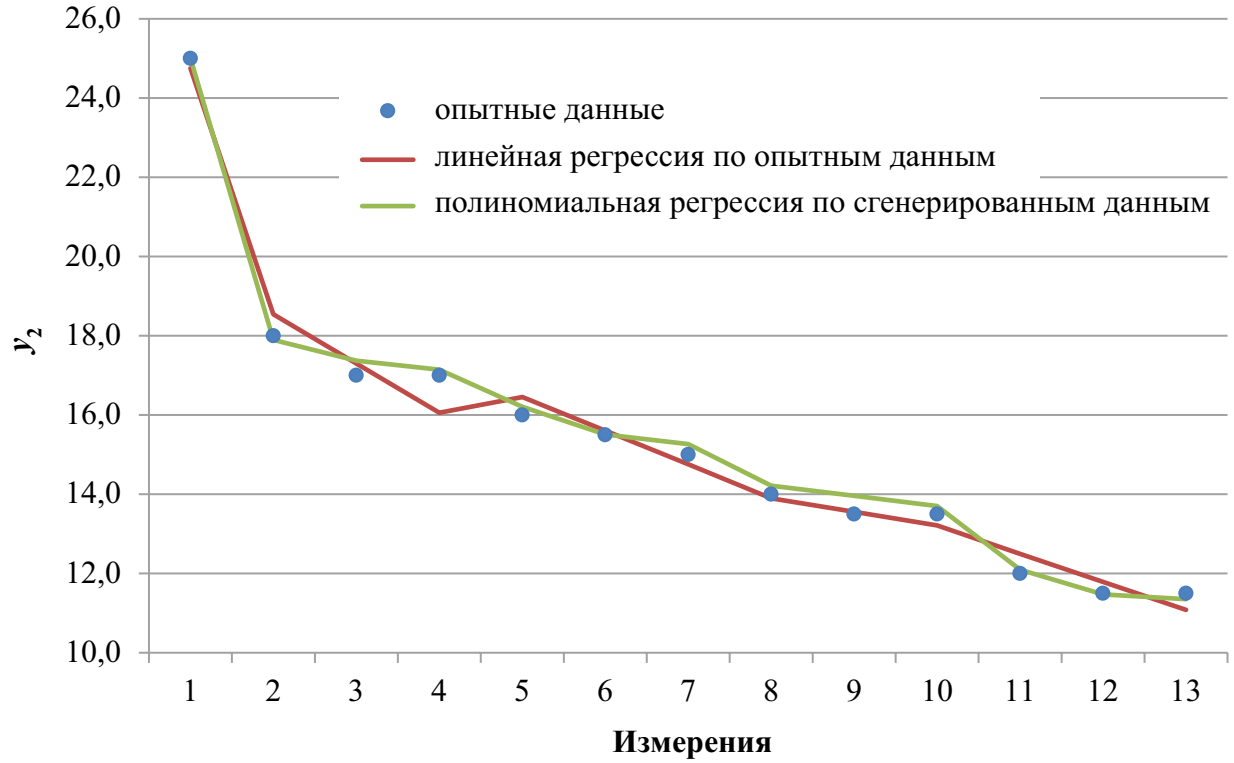

Рис. 4. Графическое представление экспериментальных и модельных значений параметра y_2

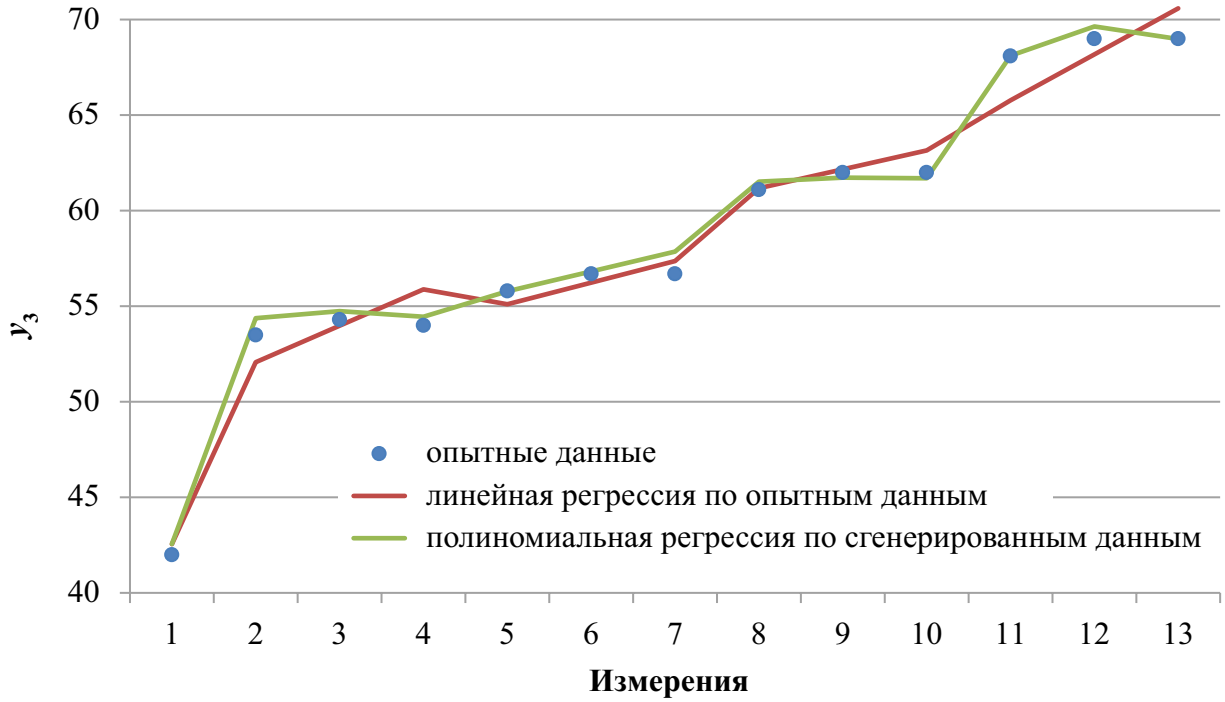

Рис. 5. Графическое представление экспериментальных и модельных значений параметра y_3

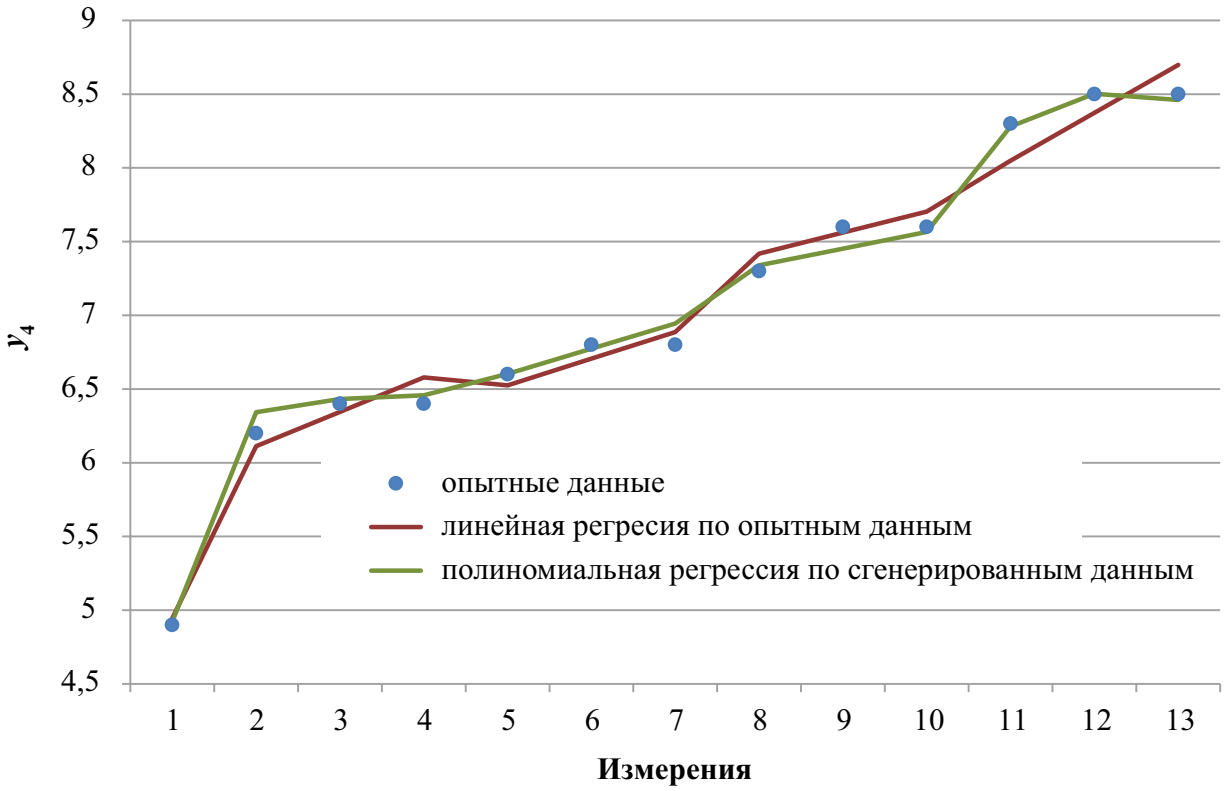

Рис. 6. Графическое представление экспериментальных и модельных значений параметра y_4

4. Увеличение объема экспериментальных данных за счет генерации их сетью CGAN, обученной на экспериментальной выборке, повышает качество регрессионных моделей за счет расширения возможности выбора вида моделей на стадии их спецификации и увеличения статистической значимости оценок параметров моделей. ▲

Библиографический список

1. Generative Adversarial NetWork / I. J. Goodfellow [et al.]. URL: https://arxiv.org/abs/1406.2661 (дата обращения 03.11.2024).
2. Грачев В. В., Федотов М. В. Повышение качества обучения эталонных диагностических моделей сложных технических объектов аугментацией обучающих данных // Автоматика на транспорте. 2023. Т. 9, № 3. С. 258–273. DOI: 10.20295/2412-9186-2023-9-03-258-273. EDN VXLQLW
3. Mehdi M., Osindero S. Conditional Generative Adversarial Nets. URL: https://arxiv.org/abs/1411.1784 (дата обращения 03.11.2024).
4. Ссылка на функцию расстояния Вассерштейна в Python. URL: https://question-it.com/questions/15429235/ssylka-na-funktsiju-rasstojanija-vassershtejna-v-python (дата обращения 03.11.2024).
5. Фостер Д. Генеративное глубокое обучение. Творческий потенциал нейронных сетей. СПб.: Питер, 2020. 336 с.: ил.
6. Айвазян С. А., Мхитарян В. С. Прикладная статистика. Основы эконометрики. Т. 1. Теория вероятностей и прикладная статистика. М.: Юнити-Дана, 2001. 656 с.
7. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Исследование зависимостей // Финансы и статистика. 1985. 487 с.
8. Кремер Н. Ш., Путко Б. А. Эконометрика / под ред. Н. Ш. Кремера. М.: Юнити-Дана, 2010. 328 с.
9. Чалганова А. А. Построение множественной регрессии и оценка качества модели с использованием табличного процессора Excel. СПб.: РГГМУ, 2022. 89 с.
10. Орлов А. И. О Эконометрика: учебник для вузов. Ростов н/Д.: Феникс, 2009. 412 с.
11. Трусова А. Ю. Анализ данных. Многомерные статистические методы: учебное пособие. Самара: Издательство Самарского университета, 2023. 92 с.
12. Генератор случайных чисел Excel в функциях и анализе данных. URL: https://exceltable.com/funkcii-excel/generator-sluchaynyh-chisel (дата обращения 03.11.2024).
13. Расстояние Махалонобиса. URL: https://habr.com/ru/articles/555144/ (дата обращения 03.11.2024).

TRANSPORT AUTOMATION RESEARCH. 2024. Vol. 10, no. 4. P. 382–394
DOI: 10.20295/2412-9186-2024-10-04-382-394

On the possibility of increasing the statistical significance of regression dependencies in the problems of processing small-volume experimental samples

Information about authors
Grachev V. V., Doctor in Engineering, Professor[1]. E-mail: v_grach@mail.ru
Schwartz M. A., PhD in Engineering, Associate Professor[2].
E-mail: shvarts4545@mai.ru
Grishchenko A. V., Doctor in Engineering, Professor[1]. E-mail: sanklok@mail.ru
Schwartz F. M., Engineer, Master[3]. E-mail: films@mail.ru

[1] Emperor Alexander I St. Petersburg State Transport University, Department of Locomotives and Locomotive Facilities
[2] Emperor Alexander I St. Petersburg State Transport University, Department of Higher Mathematics[2],
[3] Emperor Alexander I St. Petersburg State Transport University, Department of Chemistry[3]

Abstract: The article considers the problem of increasing the reliability and statistical significance of regression models of research objects built on small experimental data samples. Insufficient amount of experimental data forces the researcher to use linear models with a minimum number of variable factors, however, even with such a choice of the type of model, insufficient statistical significance of parameter estimates excludes the possibility of using it for reliable forecasting of changes in the explained variables. In order to expand the possibility of choosing the type of model at the specification stage and to increase the statistical significance of its parameter estimates, it is proposed to expand the volume of experimental data using a statistical model of the object of study, built on the basis of a generative adversarial neural network. When training on a small sample obtained during an experimental study of the object, the generator of a conditional generative adversarial network generates data clusters with centroids corresponding to the points of the training (experimental) sample. The results of the analysis of the data of a physical experiment are presented, confirming its main provisions.

Keywords: experimental sample, linear regression model, statistical significance, regression parameter estimation, conditional generative adversarial network, statistical model, multiple correlation, Euclidean – Mahalanobis distance.

References

1. Generative Adversarial NetWork / I. J. Goodfellow. URL: https://arxiv.org/abs/1406.2661 (data obrashcheniya 03.11.2024).
2. Grachev V. V., Fedotov M. V. Povyshenie kachestva obucheniya etalonnyh diagnosticheskih modelej slozhnyh tekhnicheskih ob"ektov augmentaciej obuchayushchih dannyh // Avtomatika na transporte. 2023. T. 9, no. 3. S. 258–273. DOI: 10.20295/2412-9186-2023-9-03-258-273. EDN VXLQLW (In Russian)

3. Mehdi M., Osindero S. Conditional Generative Adversarial Nets. URL: https://arxiv.org/abs/1411.1784 (data obrashcheniya 03.11.2024).
4. Ssylka na funkciyu rasstoyaniya Vassershtejna v Python. URL: https://question-it.com/questions/15429235/ssylka-na-funktsiju-rasstojanija-vassershtejna-v-python (data obrashcheniya 03.11.2024). (In Russian)
5. Foster D. Generativnoe glubokoe obuchenie. Tvorcheskij potencial nejronnyh setej. SPb.: Piter, 2020. 336 s.: il. (In Russian)
6. Ajvazyan S. A., Mhitaryan V. S. Prikladnaya statistika. Osnovy ekonometriki. T. 1. Teoriya veroyatnostej i prikladnaya statistikaj. M.: Yuniti-Dana, 2001. 656 s. (In Russian)
7. Ajvazyan S. A., Enyukov I. S., Meshalkin L. D. Prikladnaya statistika. Issledovanie zavisimostej // Finansy i statistika. 1985. 487 s. (In Russian)
8. Kremer N. Sh., Putko B. A. Ekonometrika / pod red. N. Sh. Kremera. M.: Yuniti-Dana, 2010. 328 s. (In Russian)
9. Chalganova A. A. Postroenie mnozhestvennoj regressii i ocenka kachestva modeli s ispol'zovaniem tablichnogo processora Excel. SPb.: RGGMU, 2022. 89 s. (In Russian)
10. Orlov A. I. O Ekonometrika: uchebnik dlya vuzov. Rostov n/D.: Feniks, 2009. 412 s. (In Russian)
11. Trusova A. Yu. Analiz dannyh. Mnogomernye statisticheskie metody: uchebnoe posobie. Samara: Izdatel'stvo Samarskogo universiteta, 2023. 92 s. (In Russian)
12. Generator sluchajnyh chisel Excel v funkciyah i analize dannyh. URL: https://exceltable.com/funkcii-excel/generator-sluchaynyh-chisel (data obrashcheniya 03.11.2024). (In Russian)
13. Rasstoyanie Mahalonobisa. URL: https://habr.com/ru/articles/555144/ (data obrashcheniya 03.11.2024). (In Russian)