УДК004.67

ПРИМЕНЕНИЕ КОРРЕЛЯЦИОННОГО АНАЛИЗА ДЛЯ ИССЛЕДОВАНИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Новиков С. П., Зайцева Е. Ю.
Донской государственный технический университет, Ростов-на-Дону, Российская Федерация
n_serg7@mail.ru
katerinazaytseva77@gmail.com

В статье рассматривается применение корреляционного анализа для исследования данных, получаемых при воздействии агрессивных сред на полупроводниковые датчики.

Ключевые слова: корреляция, коэффициент, диаграмма рассеяния, обработка данных, сигнал.

UDC 004.67

APPLICATION OF CORRELATION ANALYSIS FOR THE STUDY OF EXPERIMENTAL DATA

Novikov S. P., Zaytseva E.Yu.
Don State Technical University, Rostov-on-Don, Russian Federation
n_serg7@mail.ru
katerinazaytseva77@gmail.com

This article discusses the use of correlation analysis for the study of data obtained by the impact of aggressive environment on semiconductor sensors.

Keywords: correlation, coefficient, scatter diagram, data processing, signal.

Введение. Корректное предсказание поведения переменных в различных условиях является необходимым фактором успешного проектирования и применения датчиков и других устройств. В большинстве случаев между переменными существуют зависимости, при которых каждому значению одной величины (аргумента) соответствует не какое-то определенное значение другой величины, а множество ее возможных значений — определенное распределение. Такая зависимость называется стохастической или вероятностной [1].

Частным случаем вероятностной зависимости является корреляционная зависимость — стохастическая зависимость между случайными величинами, при которой наблюдается функциональная зависимость между значениями одной величины и средними значениями другой величины [2].

При проведении данного исследования использовался лабораторный стенд для изучения свойств полупроводниковых датчиков в агрессивных средах. Полученные таким образом данные обрабатывались при помощи корреляционного анализа.

В рамках представленной работы было необходимо выполнить пять перечисленных ниже задач.

1. Сформулировать основные понятия.
2. Определить данные для обработки.
3. Построить диаграмму рассеяния.
4. Посчитать коэффициент корреляции.
5. Определить зависимости данных на основе корреляционного анализа:

— скорость реакции — размах сигнала;
— скорость восстановления — размах сигнала;
— скорость реакции — скорость восстановления.

1. Корреляция. Диаграмма рассеяния. Корреляционный анализ основывается на предположении о том, что между количественными переменными существует линейная зависимость, позволяющая измерить степень связи между ними.

Для наглядного представления связи между переменными применяется диаграмма рассеяния (рис. 1).

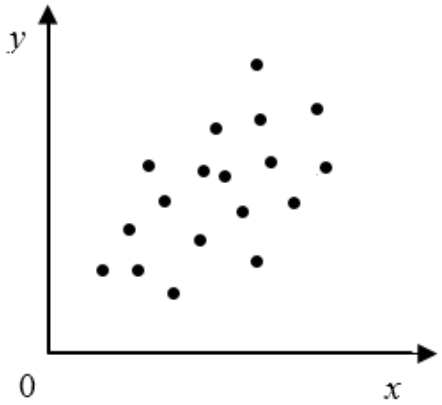

Рис. 1. Диаграмма рассеяния

Здесь x и y — значения первой и второй переменных. Образованное скопление (облако) точек определяет картину отношения между переменными. По ширине разброса точек можно сделать вывод о силе связи [3].

Параметр, который систематически увеличивается и (или) уменьшается другим, называется параметром управления (или независимой переменной) и обычно отображается на графике вдоль горизонтальной оси. Зависимая переменная обычно располагается вдоль вертикальной оси. Если никакой зависимости не существует, любой тип переменной может быть нанесен на любую ось, и график рассеяния будет иллюстрировать только степень корреляции (не причинно-следственной связи) между двумя переменными [3].

Аналитическая функция, аппроксимирующая (приближенно описывающая) наблюдаемые эмпирические значения, называется функцией регрессии (рис. 2).

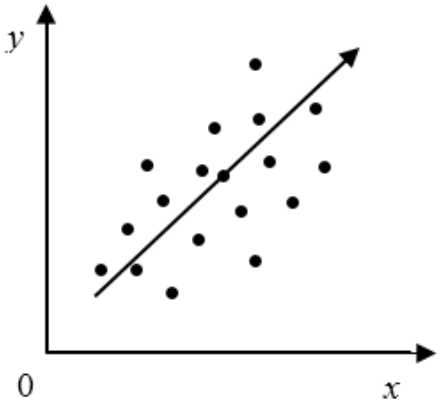

Рис. 2. Функция регрессии

Функция регрессии отражает тенденцию изменения одной величины под действием другой и строится таким образом, чтобы эмпирические точки корреляционного поля лежали как можно ближе к ней [3, 4]. Эксперт выбирает функцию регрессии (линейную, параболическую, гиперболическую или логарифмическую).

Диаграмма рассеяния и функция регрессии позволяют проводить визуальный анализ данных, а для определения количественного значения необходимо рассчитать коэффициент корреляции.

2. Коэффициент корреляции. Коэффициент корреляции показывает степень взаимосвязи между двумя изучаемыми переменными. Коэффициент корреляции r_{xy} рассчитывается по следующей формуле:

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sigma(x)\sigma(y)}, \quad (1)$$

где σ(x), σ(y) — среднеквадратическое отклонение.

Средние значения:

$$\overline{x} = \frac{\sum x_i}{n};$$

$$\overline{y} = \frac{\sum y_i}{n};$$

$$\overline{x \cdot y} = \frac{\sum x_i \cdot y_i}{n}.$$

Здесь x и y — значения переменных, n — размер выборки.

Значение коэффициента корреляции можно интерпретировать следующим образом:

— если r_{xy} равно 1, то между двумя значениями существует идеальная положительная корреляция;

— если r_{xy} равно –1, то между двумя значениями существует идеальная отрицательная корреляция;

— если r_{xy} равно нулю, то между двумя значениями нет корреляции [4].

Если коэффициент корреляции (1) равен –1 или +1, то точки на диаграмме рассеяния будут лежать точно на прямой линии, что указывает на сильную корреляцию между переменными.

При положительной корреляции увеличение (или уменьшение) значений одной переменной ведет к закономерному увеличению (или уменьшению) другой переменной.

При отрицательной корреляции увеличение (или уменьшение) значений одной переменной ведет к закономерному уменьшению (или увеличению) другой переменной.

При положительной или отрицательной корреляции точки на графике не должны располагаться на прямой линии [5, 6].

Если линейных отношений нет, то корреляция равна нулю, и точки на графике разбросаны случайным образом. Однако нелинейные отношения не обязательно обусловливают отсутствие связи между переменными. Так как невозможно однозначно сказать о связи между переменными, необходимо ввести текущий контроль данных на основании графического отображения (рис. 3).

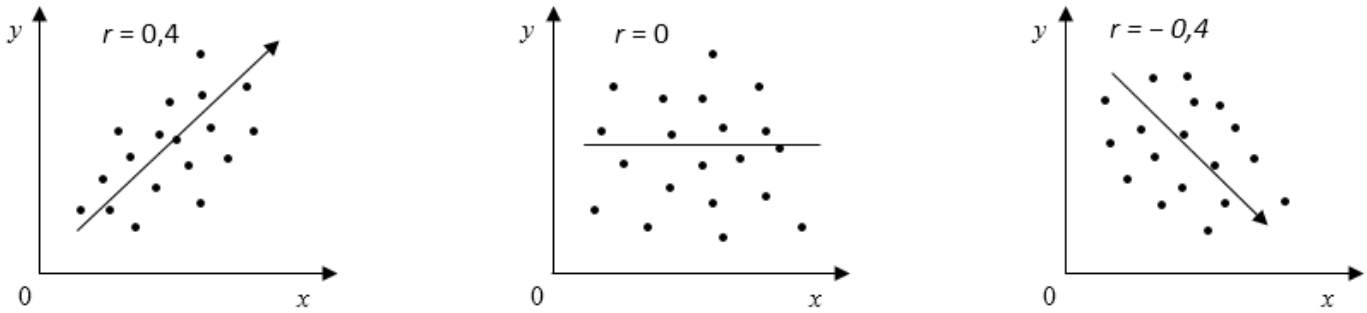

Рис. 3. Графическое отображение данных

Сильная корреляция между переменными не означает, что одна из них оказывает влияние на другую, поскольку многие переменные изменяются с течением времени и, следовательно, могут коррелировать [6].

3. Серия измерений. В качестве примера рассмотрена следующая задача. На подготовленный датчик воздействует агрессивная среда. Измеряется время реакции. Затем, после

продувки, измеряется время восстановления. Собранные данные, поступающие с лабораторного стенда, прошли фильтрацию от ложных сигналов и шумов (рис. 4).

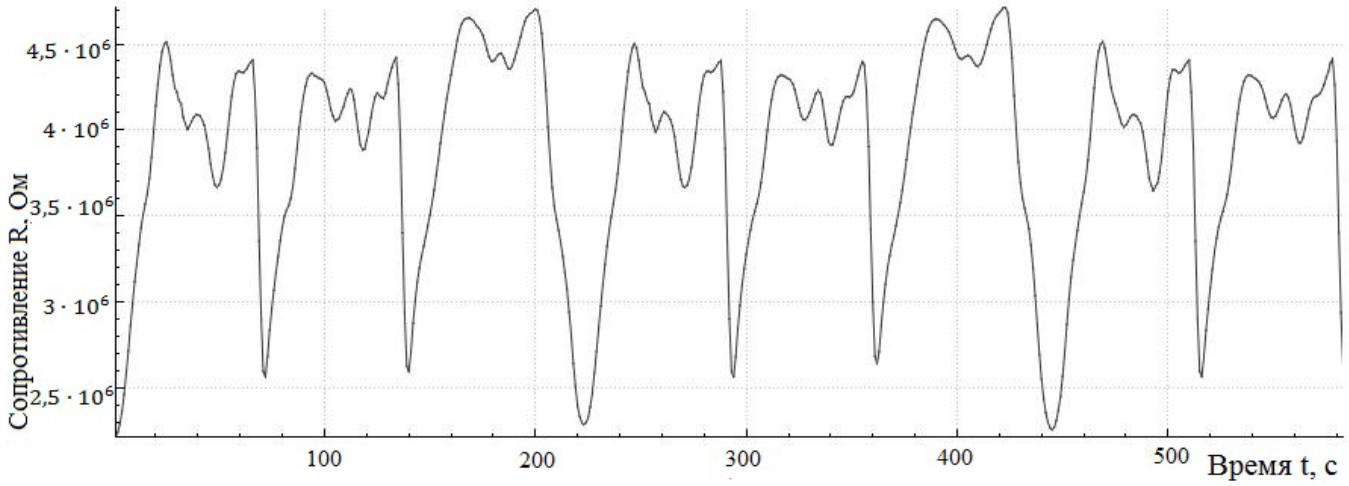

Рис. 4. График экспериментальных данных после сглаживания

Следует проанализировать обработанные данные для определения тенденции их влияния на стабильность, скорость реакции и концентрацию вещества.

3. Корреляционный анализ. Для загруженных данных (рис. 5) необходимо построить диаграмму рассеяния и посчитать коэффициент корреляции.

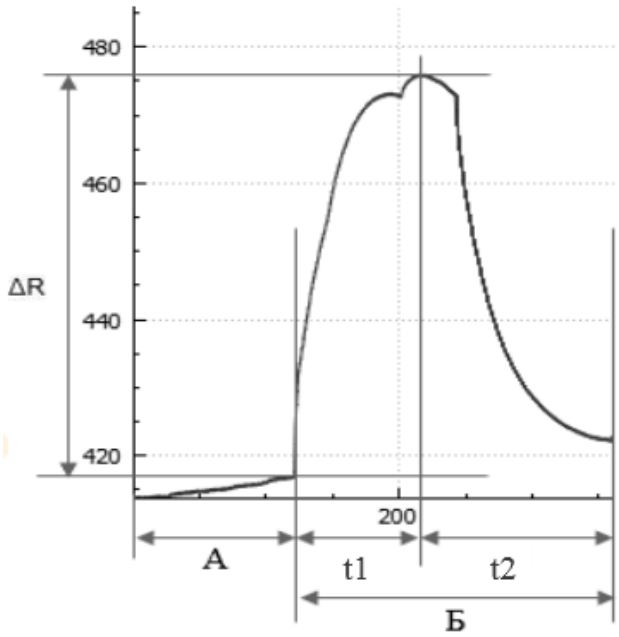

Рис. 5. Выбор параметров: А — выход в рабочий режим (заполнение камеры газом); Б — рабочий режим, т. е. срабатывание сенсора, измерение (получение данных); t1 — время реакции; t2 — время восстановления; ΔR — размах изменения сопротивления

3.1. Анализ скорости реакции и размаха сигнала. Анализ зависимости между скоростью реакции и размахом сигнала позволил получить диаграмму рассеяния (рис. 6).

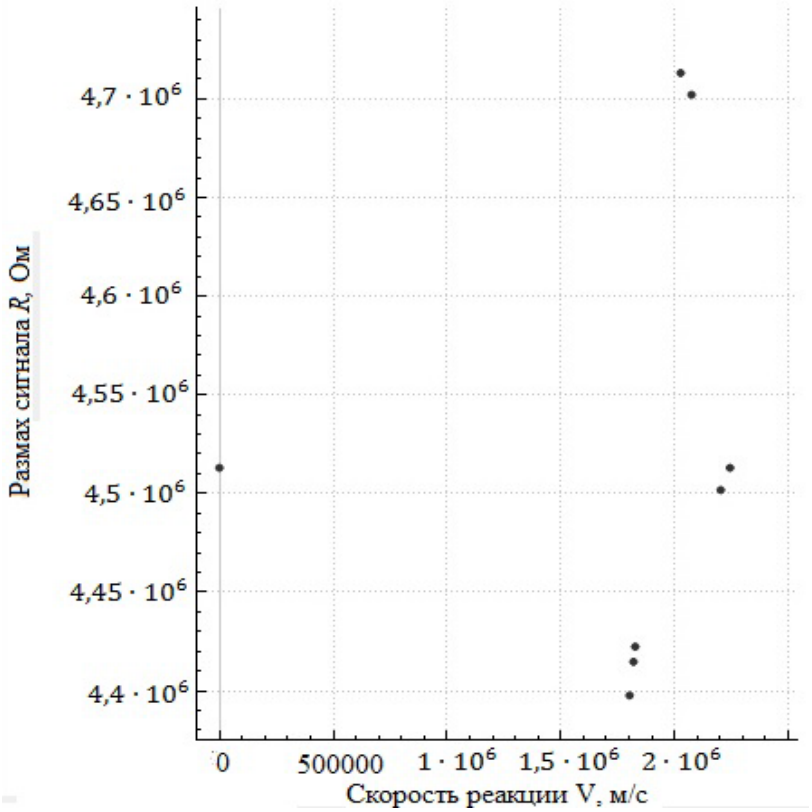

Рис. 6. Диаграмма рассеяния значений

Далее была определена и отображена линия регрессии (рис. 7).

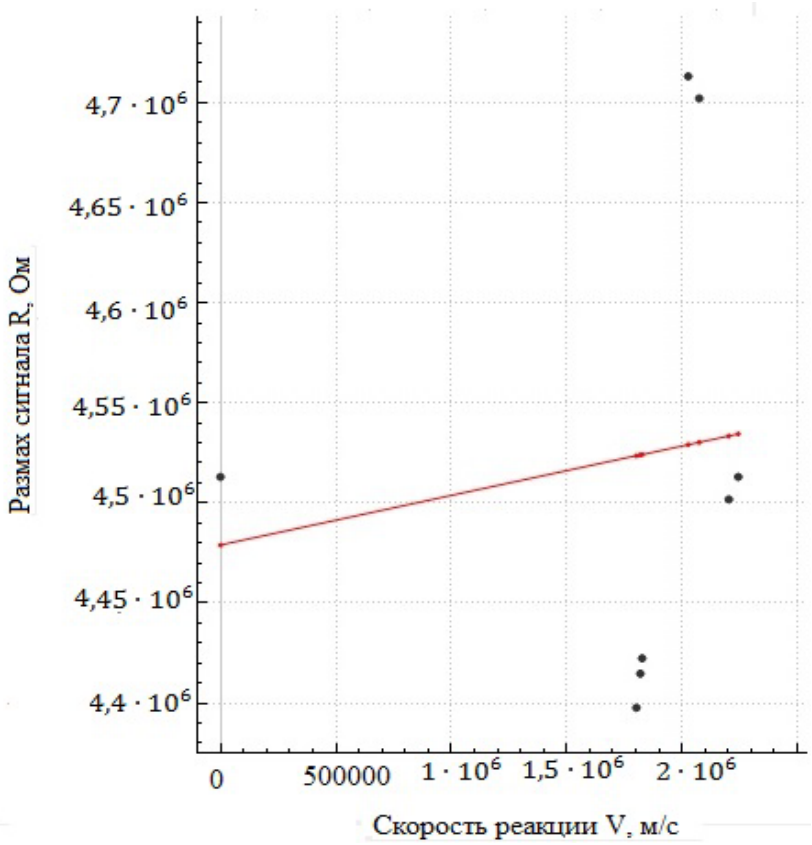

Рис. 7. Функция регрессии проводимого анализа

Отношение между рассматриваемыми данными — среднее. Об этом говорит коэффициент корреляции r = 0,52.

3.2. Анализ скорости восстановления и размаха сигнала. Связь между скоростью восстановления и размахом сигнала показана на рис. 8.

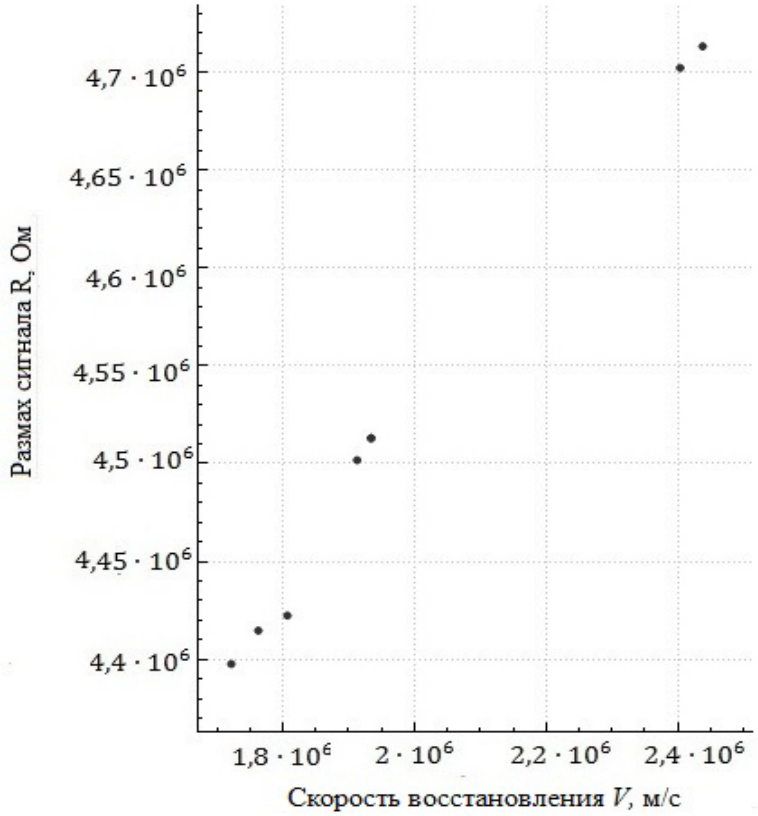

Рис. 8. Диаграмма рассеяния значений

Линия регрессии для проводимого анализа изображена на рис. 9.

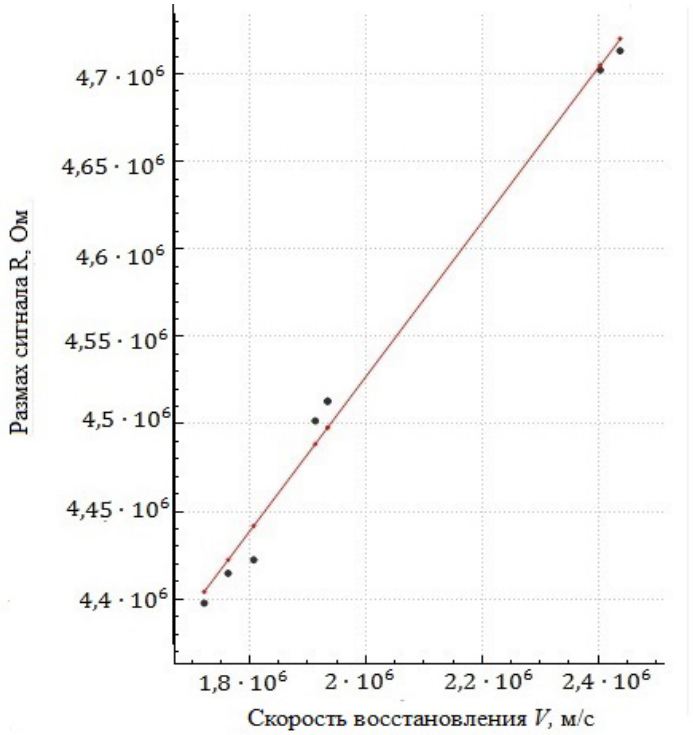

Рис. 9. Функция регрессии для проводимого анализа

Коэффициент r = 0,16, что свидетельствует о слабой корреляции между данными.

3.3. Анализ скорости реакции и скорости восстановления. На рис. 10 визуализирован результат анализа корреляции между скоростью реакции датчика и скоростью восстановления.

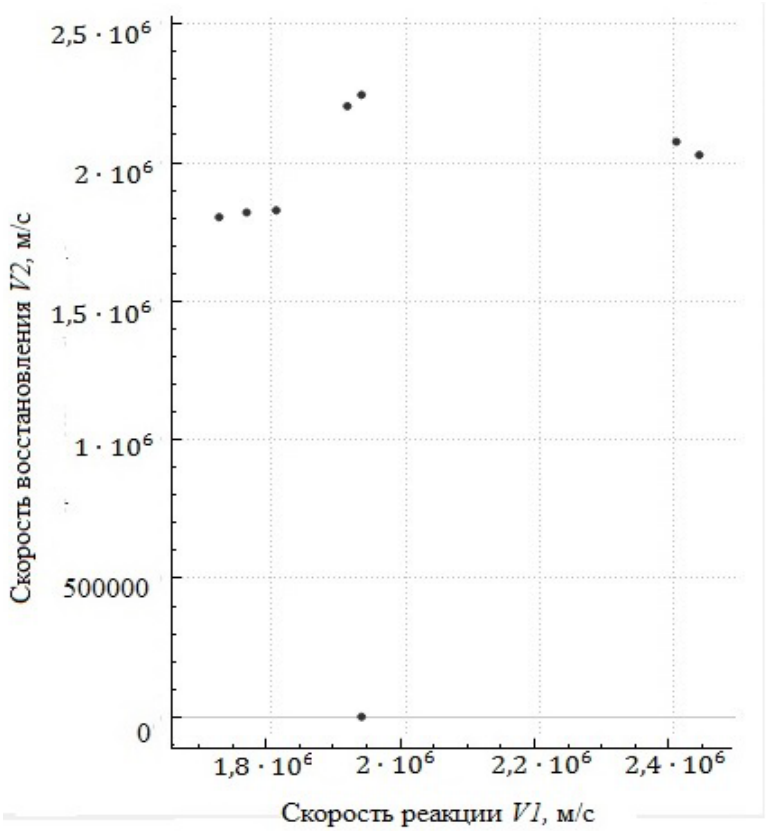

Рис. 10. Диаграмма рассеяния значений

Линия регрессии для проводимого анализа представлена на рис. 11.

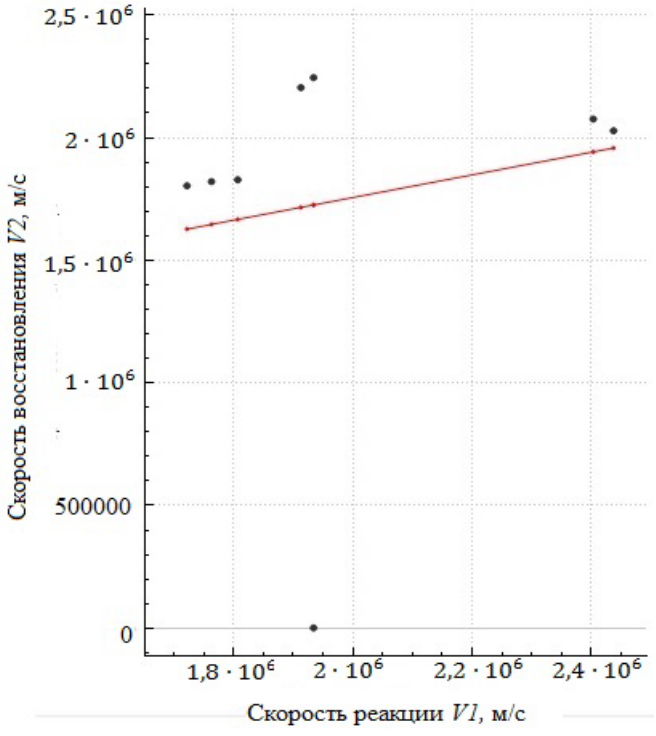

Рис. 11. Функция регрессии для проводимого анализа

Коэффициент корреляции рассматриваемых данных принимает значение r = 0,47. Это говорит о среднем отношении между скоростью реакции датчика, подвергаемого воздействию агрессивной среды, и скоростью восстановления.

Выводы. Корреляционный анализ и большой набор данных позволяют определить зависимости величин данных. По результатам проведенного анализа можно судить о силе влияния скоростных характеристик датчика на стабильность, скорость реакции и концентрацию вещества. Для более точных результатов необходимо провести дополнительное исследование тестируемых образцов.

Библиографический список

1. Общая теория статистики / Под ред. Р. А. Шмойловой. — 3-е изд., перераб. — Москва : Финансы и статистика, 2002. — 656 с.

2. Фёрстер, Э. Методы корреляционного и регрессионного анализа. Руководство для экономистов / Э. Фёрстер, Б. Рёнц. — Москва : Финансы и статистика, 1983. — 304 с.

3. Гмурман, В. Е. Теория вероятностей и математическая статистика / В. Е. Гмурман. — 10-е изд., стереотип. — Москва : Высшая школа, 2004. — 479 с.

4. Гржибовский, Л. М. Анализ порядковых данных / А. М. Гржибовский // Экология человека. — 2008. — № 11. — С. 48–55.

5. Елисеева, И. И. Общая теория статистики / И. И. Елисеева, М. М. Юзбашев. — 4-е изд., перераб. и дополн. — Москва : Финансы и статистика, 2002. — 480 с.

6. Благовещенский, Ю. Н. Тайны корреляционных связей в статистике / Ю. Н. Благовещенский. — Москва : Научная книга ; Инфра-М, 2009. — 156 с.