

АНАЛИЗ СТАТИСТИЧЕСКИХ ДАННЫХ ОБ ОТКАЗАХ ПРОМЫСЛОВЫХ ТРУБОПРОВОДОВ

Д.П. Кармачев

(г. Томск, Томский политехнический университет)

E-mail: karmachevd@mail.ru

FAILURE STATISTICS DATA ANALYSIS OF PIPELINES

D.P. Karmachev

(Tomsk, Tomsk Polytechnic University)

Abstract. The paper presents an exploratory analysis of statistical data on failures. In the process of analysis, each of the continuous and categorical features of the original sample is examined in detail, and steps to eliminate some of the features are described. The author presents tables of the various continuous features importance. The result of the work is the final prepared data table for use in problems of predicting pipeline failures for internal corrosion reasons.

Keywords: data analysis, pipeline failures, failure prediction, data reduction, importance of features.

Введение. В процессах первичной подготовки нефти, газа и воды применяются различные по конструктивным характеристикам промышленные трубопроводы (ПТ), которые эксплуатируются под воздействием различных внутренних и внешних факторов. При этом для отказов ПТ характерны высокие материально-экономические потери, а также негативное влияние на экологическую обстановку. В связи с этим, нефтяными компаниями регулярно выполняются работы по созданию новых и модернизации действующих систем поддержки принятия решения (СППР) при эксплуатации промышленных трубопроводов. Для подсистемы прогнозирования (в составе СППР) исходными данными являются не только динамические модели, но и статистические данные об отказах промышленных трубопроводов. [1] Перед применением накопленной исторической информации в прогнозных моделях важным шагом является проведение фильтрации исходной выборки и разведочного анализа данных. Целью данной работы является проведение разведочного анализа статистических данных об отказах промышленных трубопроводов определенной группы месторождений одной из нефтяных компаний РФ.

Анализ исходной выборки. Исходная выборка была получена путём выгрузки необходимых данных из информационной системы OisPipe, охватывающей определенную группу нефтяных месторождений одной из нефтяных компаний, действующих на территории центральной России. Исходная выборка охватывает отказы промышленных трубопроводов за период 2000-2017 гг. Признаки ПТ исходной выборки представлены в таблицах 1, 2. Общая размерность исходной выборки составляет 15143 строки.

На первом этапе произведена сравнительная оценка количества отказов произошедших по различным причинам. К основным причинам отказов ПТ относятся: внутренняя коррозия, дефект сварки, заводской брак, механические повреждения, внешняя коррозия, повышение давления, нарушение правил эксплуатации, конструктивный недостаток, строительный брак и другие. В исходной выборке среди других признаков в подавляющем случае обладает информация касательно характеристик участка ПТ (длина, диаметр, толщина стенки), а также информация касательно параметров перекачки и характеристик перекачиваемых сред. Также важно отметить, что около 70% всей выборки (10634 строки) – это отказы, произошедшие по причине внутренней коррозии. Обобщая всю данную информацию, в рамках данного предварительного анализа было принято решение сократить исходную выборку и продолжить работу только с информацией об отказах, произошедших по причине внутренней коррозии.

Таблица 1. Перечень категориальных признаков исходной выборки данных об отказах

Наименование	Преобладающие категории	Размерность
Наименование месторождения	-	13
Наименование цеха и площадки	-	71
Назначение участка	выкидные линии скважин (5549); нефтеборные сети (2862); высоконапорные водоводы (2427); низконапорные водоводы (1191); нагнетательные линии скважин (1160); напорные нефтепроводы (673); внутриплощадочные водоводы (509); внутриплощадочные нефтепроводы (259).	20
Материал промышленного трубопровода	сталь 20 (9053); гибкие полимерно-металлические (ГПМ) (1761); сталь 10 (1692).	24
Тип внутренней изоляции	нет покрытия (9644); полиэтилен без марки (2132); полиэтилен высокого давления 153-10К ГОСТ 16336-77 (281); полиэтилен низкого давления (161); эпоксидное покрытие (118).	19
Тип внешней изоляции	битум различных марок (6308); полилен различных марок (621); праймер (762); полимерное покрытие (551); полиэтилен (275);	52
Завод изготовитель	-	49

Таблица 2. Перечень непрерывных признаков исходной выборки данных об отказах

Наименование	Диапазон
Длина, м	0.4 – 38000
Диаметр, мм	16 – 1420
Толщина стенки, мм	3 – 30
Рабочее давление, МПа	0.01 – 27
Давление в момент отказа, МПа	0.01 – 28
Расход жидкости, м ³ /сутки	0.01 – 60000
Расход нефти, т/сутки	0.03 – 19850
Газовый фактор, м ³	0.01 – 110
Обводненность (процентное содержание воды в среде), %	0 – 100
Скорость потока, м/с	0.001 – 33
Температура потока, С°	0,05 – 100

На втором этапе анализа рассматриваются непрерывные и категориальные признаки, описывающие материалы ПТ, типы и материалы внутренних покрытий ПТ, а также габаритные характеристики ПТ.

В рамках текущего исследования толщина стенки и диаметр участков рассматриваются как непрерывные признаки, при том, что по отношению к каждому конкретному примеру исходной выборки соответствуют определенные нормированные значения данных призна-

ков. Некоторые примеры типов диаметров – 95 мм, 114 мм, 168 мм, 273 мм, 325 мм и т.д., а также типов толщин стенок – 4 мм, 4.5 мм, 5 мм, 5.5 мм, 6 мм, 7 мм, 8 мм и т.д. Рассмотрение диаметра, как непрерывного признака, обусловлено наличием зависимостей между параметрами перекачки, габаритными характеристиками, типами течений сред, а также типами структур газожидкостных потоков. При этом в дальнейших исследованиях предполагается установить явные корреляции между типами течений, типами структур и габаритными характеристиками, параметры которых определены не нормированными значениями, а непрерывными. Известно, что динамика развития коррозионных и эрозионных процессов обратно пропорциональна толщине стенки [2], дальнейшие исследования направлены на нахождение данных зависимостей не только по отношению к нормированным значениям толщин, но также и по отношению к промежуточным значениям, означающим определенную степень износа участка ПТ, поэтому толщина стенки рассматривается как непрерывный признак. Несмотря на то, что габаритные характеристики рассматриваются в качестве непрерывных признаков, на следующем этапе предварительной обработки были исключены строки данных, описывающие отказы нетипичных участков ПТ по толщине стенки и диаметру. Причиной исключения является недостаточное количество примеров отказов (до 5) по отношению к каждому из значений данных признаков. Исключены участки, имеющие следующие значения диаметра – 57 мм, 60 мм, 70 мм, 80 мм, 102 мм, 105 мм, 1020 мм, 1400 мм; и следующие значения толщины стенки – 3 мм, 3.5 мм.

Материал ПТ рассматривается в качестве категориального признака. В рамках разведочного анализа из исходной выборки исключены примеры отказов участков ПТ, изготовленных из стали 3 серии, а также из специфических полимеров, по причине их недостаточного количества (до 15) по отношению к каждому из типов материалов ПТ. Размерность данного признака сократилась с 24 до 13 по отношению к исходной выборке.

Категориальный признак «Тип внешней изоляции» также был полностью исключен из исходной выборки, так как в рамках планируемых исследований будут рассматриваться отказы, произошедшие по причине внутренней коррозии.

В исходной выборке представлены участки ПТ с 12-ти различными типами внутренней изоляции. При этом 4363 примеров отказов связаны с ПТ, эксплуатируемыми без внутреннего покрытия. Наличие внутренней изоляции значительно снижает аварийность по причинам внутренних коррозий [3], за счет устранения непосредственного контакта воды, содержащейся в перекачиваемой среде, с металлическим телом трубы. В исходной выборке присутствует 441 пример отказов ПТ, тип внутренней изоляции которых неизвестен. В таблице дано примечание «Нет данных». Данные примеры отказов исключаются в рамках настоящего исследования для того, чтобы за ранее нивелировать возможные неточности в будущих исследованиях, так как предполагается, что данный признак является достаточно весомым в рамках решения задачи прогнозирования сроков службы ПТ.

На третьем этапе анализа были рассмотрены основные непрерывные признаки, определяющие условия эксплуатации участков промысловых трубопроводов – параметры перекачки и свойства перекачиваемых сред.

Наличие воды в перекачиваемой среде определяет площадь соприкосновения жидкой фазы, следовательно, напрямую влияет на возникновение и динамику протекания окислительных процессов. [4] Исходная выборка содержит 752 примера отказов различных ПТ с пропущенными значениями по обводненности. Для всех пропущенных значений по обводненности, но относительно водоводов высокого и низкого давления, были восстановлены значения в 100%, условно означающие абсолютное содержание воды. [2] Примеры отказов нефтесборов и газоконденсатопроводов с пропущенными значениями по обводненности были исключены из исходной выборки, восстановить пропущенные значения в данном случае не представляется возможным.

Газовый фактор косвенно определяет количество углекислого газа в перекачиваемой среде, который в свою очередь влияет на динамику образования и протекания коррозионных процессов. [2] Для водоводов были восстановлены пропущенные значения газового фактора

равного в среднем 1 м^3 . Данное значение обусловлено тем, что содержание газа в воде стремится к нулю, что обусловлено спецификой процессов сепарации. [5]

Скорость потока перекачиваемой среды зависит от габаритных характеристик промысловых трубопроводов, а также от таких параметров перекачки, как давление и расход. [6] Исходная выборка содержала 8345 примеров отказов ПТ, скорость потока перекачиваемых сред в которых неизвестна. Учитывая данную информацию, а также зависимости, описанные выше, было принято решение об исключении из дальнейшего анализа данного признака.

Известно, что повышение температуры ускоряет анодные и катодные процессы за счет увеличения скорости движения ионов, что напрямую влияет на образование коррозий и динамики развития коррозионных процессов в целом. [3] При этом в дальнейших исследованиях данным признаком пришлось пренебречь, в связи с тем, что в 6782 случаях неизвестна информация о значениях температур транспортируемых сред.

Анализ статистической информации проводится с целью выявления зависимостей в прогнозировании отказов по причинам коррозий, которые являются результатом продолжительных процессов транспортировки определенных сред в рамках усредненных параметров перекачки, поэтому признак «давление в момент отказа» не рассматривается в рамках дальнейших исследований. Также необходимо отметить, что планируемая точность прогнозирования отказов будет составлять 1-6 месяцев, в связи с чем, применение признака «давление в момент отказа» в дальнейших исследованиях нецелесообразно.

Результаты анализа. Сводная информация касательно итоговой выборки статистических данных об отказах по результатам проведенного разведочного анализа представлена в таблице 3.

Таблица 3. Сводная информация по итоговой выборке данных об отказах

Наименование	Диапазон / Размерность	Тип признака
Наименование месторождения	13	Категориальный
Наименование цеха и площадки	67	Категориальный
Материал трубы	7	Категориальный
Тип внутренней изоляции	13	Категориальный
Длина, м	0.6 – 13600	Непрерывный
Диаметр, мм	73 – 1420	Непрерывный
Толщина стенки, мм	4 – 16	Непрерывный
Рабочее давление, МПа	0.01 – 19	Непрерывный
Расход жидкости, $\text{м}^3/\text{сутки}$	0.01 – 60000	Непрерывный
Расход нефти, т/сутки	0.03 – 19850	Непрерывный
Газовый фактор, м^3	0.01 – 110	Непрерывный
Обводненность, %	0 – 100	Непрерывный
Время наработки на отказ, месяц	3563 – 28774	Непрерывный

Общая размерность подготовленной выборки составляет 5587 строк. В рамках разведочного анализа не выявлены явные зависимости между временем наработки на отказ и отдельно взятыми непрерывными признаками.

Подготовленная выборка, состоящая из 5587 строк, была разбита случайным образом в соотношении 70% к 30% на тренировочный и тестовый наборы данных соответственно. Предварительная оценка подготовленных данных была выполнена в ППП Python с применением ансамблевого алгоритма случайный лес (random forest). [8] [9].

Коэффициент детерминации R^2 [7] для тестового набора данных составил 0.942, для тренировочного 0.678.

В процессе проводимого разведочного анализа были составлены дополнительные точечные выборки относительно материалов труб и типов внутренней изоляции. На основе составленных выборок, учитывающих, в том числе признак «Скорость потока», была выполнена оценка важности непрерывных признаков. Оценка выполнялась в ППП Python с помощью классификатора ExtraTreesClassifier. Также для каждой точечной выборки определены коэффициенты детерминации R^2 . Результаты оценки представлены в таблице 4.

По результатам оценки видно, что скорость потока оказывает существенное влияние на аварийность промышленных трубопроводов. При этом из исходной выборки данный признак был исключен, так как в 8345 примерах отказов значения скорости потока неизвестны. Дальнейшие исследования сходимости данных и оценка точности прогнозируемых величин позволят определить более детально степень важности данного признака при условии наличия альтернативной информации о параметрах перекачки (расход, давление).

Динамика распределения показателей важности признаков для водоводов напрямую связана с наличием, либо отсутствием внутреннего покрытия. ГПМ трубы имеют наиболее низкий показатель удельной аварийности в исходной выборке, также, по результатам оценки важности признаков видно, что аварийность в большей степени данных типов труб связана с количеством перекачиваемой водонефтяной эмульсии и скоростью ее перекачки. На основе результатов, можно предположить, что коррозии в данных типах труб развиваются с наименьшей интенсивностью в сравнении с ПТ марки стали 20 и 10, эксплуатируемых без применения внутреннего покрытия. Динамика распределения важности признаков в случае с ГПМ аналогична динамике в случае с водоводами, эксплуатируемыми с применением внутренних покрытий, за исключением того, что расход перекачиваемой воды в большей степени влияет на аварийность водоводов, чем скорость потока.

Таблица 4. Важность непрерывных признаков в рамках точечных выборок

Назначение	Материал, тип покрытия	R ²	Признак	Важность
Нефтебурсы	Сталь 20, нет покрытия	0.707	Скорость потока	0.239
			Обводненность	0.222
			Давление	0.206
			Расход	0.189
			Газовый фактор	0.144
	Сталь 10, нет покрытия	0.698	Скорость потока	0.309
			Расход	0.213
			Обводненность	0.201
			Давление	0.148
	ГПМ	0.739	Газовый фактор	0.129
			Скорость потока	0.433
			Расход	0.400
			Давление	0.072
			Обводненность	0.063
			Газовый фактор	0.032
Водоводы	Сталь 20, нет покрытия	0.671	Скорость потока	0.581
			Расход	0.274
			Рабочее давление	0.145
	Сталь 20, ПВД 153-10К	0.686	Расход	0.438
			Рабочее давление	0.375
			Скорость потока	0.188
	Сталь 20, полиэтилен без марки	0.634	Расход	0.496
			Рабочее давление	0.298
			Скорость потока	0.206

Заключение. По результатам проведенного разведочного анализа была сформирована итоговая выборка об отказах промышленных трубопроводов определенной группы месторождений, а также проведена оценка важности непрерывных признаков в рамках точечных выборок. Полученная информация будет использована в математических моделях в рамках прогнозирования отказов ПТ, произошедших по причинам внутренних коррозий, с применением методов машинного обучения. В исходной выборке отсутствовала информация о физико-химических свойствах перекачиваемых сред, которые оказывают существенное влияние на образование коррозий. Косвенно данные зависимости будут прослеживаться за счет наличие информации о месторождениях и площадках, физико-химические свойства сред на которых приблизительно равны в рамках определенного объекта. [2]

ЛИТЕРАТУРА

1. А.В. Аржиловский, А.В. Алферов, Р.И. Валиахметов. Концепция системы мониторинг надежности и эксплуатации промысловых трубопроводов // Нефтяное хозяйство – сентябрь 2018, С. 128-132.
2. А.И. Владимиров, В.Я. Кершенбаума. Промышленная безопасность и надежность магистральных трубопроводов: учебник для вузов – М.: Изд-во Национального институт нефти и газа, 2009. – 696 с.
3. А.В. Рудаченко, С.С. Байкин. Эксплуатационная надежность трубопроводных систем: учебное пособие для вузов – Томск: Изд-во Томского политехнического университета, 2008 – 119 с.
4. Р.И. Тавастшерна. Изготовление и монтаж технологических трубопроводов: учебное пособие для вузов – Москва: Изд-во «Высшая школа», 1967 – 282 с.
5. К.Р. Низамов. Повышение эксплуатационной надежности промысловых трубопроводов: диссертационная работа на соискание степени доктора технических наук, специальность ВАК РВ 25.00.17 – Уфа, 2001 – 300 с.
6. С.А. Ахметов. Технология глубокой переработки нефти и газа: учебное пособие для вузов — Уфа: Гилем, 2002 – 672 с.
7. Бахрушин В.Е. Методы оценивания характеристик нелинейных статистических связей // Системные технологии. – 2011 – №2(73) – С. 9-14.
8. Leo Breiman. Random Forest // Machine Learning (journal): journal. – 2001 – Vol.45, no 1 – P. 5-32.
9. Hastie T., Tibshirani R., Friedman J. Chapter 15. Random Forest // The Elements of Statistical Learning: Data Mining, Inference, and Prediction – 2nd ed. – Springer-Verlag, 2009 – 746p.

ИЗВЛЕЧЕНИЕ ЗНАЧИМЫХ ДИАГНОСТИЧЕСКИХ КРИТЕРИЕВ ДЛЯ РОЖИСТЫХ ВОСПАЛЕНИЙ ИЗ ИСТОРИИ БОЛЕЗНИ

Е.В. Кащеева

(г. Томск, Томский политехнический университет)

e-mail: ev.kashcheeva@mail.ru

EXTRACTION OF SIGNIFICANT DIAGNOSTIC CRITERIA FOR ERYSIPELAS FROM THE HISTORY OF THE DISEASE

E.V. Kashcheeva

(Tomsk, Tomsk Polytechnic University)

Abstract. This article describes the process of extracting significant diagnostic criteria for erysipelas from the history of the disease. The format of documents generated by medical information systems is inconvenient for data analysis. This article highlights the main stages of translating information about patients' complaints into a convenient form for analysis. Search for the presence or absence of patient complaints is carried out by keywords. The created program is a universal tool for the allocation of certain data about patients from the medical history in a convenient form for analysis.

Keywords: history of the disease, significant criteria, keyword search, erysipelas.

Введение. В настоящее время существует большое количество медицинских информационных систем, предназначенных для ввода данных о пациентах, а также формирования отчетов и документов на естественном языке. Сформированные документы имеют определенную структуру, разделены на блоки. Однако, для анализа данных подобной структурированности не достаточно, формат представления данных не является удобным.

Такая совокупность документов как «История болезни» помимо прочего включает в себя документ «Осмотр лечащим врачом». В данный документ вносятся сведения, получен-