УДК 550.8.05:681.3

Р.И. Ивановский, М.А. Новожилов

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ ГЕОФИЗИЧЕСКОГО МОНИТОРИНГА

Во многих прикладных задачах геофизики, сейсмологии и вулканологии, связанных с оценкой возможности землетрясений и извержения вулканов, изучением состава залегающих пород, используются группы территориально распределенных сенсоров, мониторинг которых позволяет оценить текущее состояние земной коры и динамику его изменения [1, 2]. Задачи подобного типа продолжают оставаться актуальными, поскольку от их успешного решения зависят благополучие и безопасность общества.

Необходимость обработки данных группы датчиков возникает также и в других областях, например, в энергетике, тензометрии, медицине (в части энцефалографии).

Задачи такого типа имеют ряд общих черт: многомерность получаемых данных, необходимость их совместной обработки, высокие требования к объективности оценок и прогноза. Все это, в свою очередь, определяется качеством используемых подходов, выбором технологии обработки данных мониторинга. По мнению авторов, ресурсы повышения качества решения таких задач еще не исчерпаны, дополнительные резервы могут быть найдены с применением специальных разделов математической статистики (регрессионного и корреляционного анализа) [3, 4], теории случайных процессов [4] и динамических систем.

Рассмотрим принципы формирования технологии обработки данных мониторинга с использованием указанных методов. Изложение будем сопровождать иллюстрациями в виде копий фрагментов Mathcad-файлов.

Пусть в выбранном регионе используется совокупность m однотипных датчиков, территориально распределенных по поверхности Земли или заглубленных. Измеряются, например, колебания земной коры в местах расположения датчиков; периодически осуществляется мониторинг.

В качестве варианта примем, что каждый k-й однократный мониторинг проводится на интервале времени t_D; данные каждого датчика фиксируются синхронно, на определенной частоте, последовательностями по n значений в каждой. Таким образом, в результате k-го однократного мониторинга получается массив данных, который имеет вид $(n \times m)$-матрицы $\mathbf{M}_k$.

Длительность интервала t_D и частота съема данных могут выбираться в широком диапазоне, с учетом лишь требования необходимой длины выборки, например, $n > 100$. Часто применяют $t_D = 1$ мин, при частоте более 200 гц, что с запасом удовлетворяет этому требованию.

Представление исходных данных в виде матриц $\mathbf{M}_k$ удобно, поскольку позволяет легко переходить от обработки всего массива на полном временном интервале t_D к обработке данных на коротких временных отрезках, используя для этого лишь соответствующий набор строк матрицы $\mathbf{M}_k$. На рис. 1 приведена типовая матрица $\mathbf{M}_k$ с выделенной частью, подлежащей текущей обработке.

-0,23	5,88	3,10	0,26	3,05
0,10	-2,60	-1,37	-0,12	-1,34
-0,07	1,70	0,90	0,08	0,88
0,18	-4,48	-2,36	-0,20	-2,32
•••				
-0,06	1,49	0,79	0,07	0,77
-0,13	3,34	1,76	0,15	1,73
-0,17	4,28	2,26	0,19	2,22
0,10	-2,63	-1,39	-0,12	-1,36
0,03	-0,86	-0,46	-0,04	-0,45
0,09	-2,23	-1,18	-0,10	-1,16
0,04	-0,93	-0,49	-0,04	-0,48

Рис. 1. Типовая матрица $\mathbf{M}_k$

Решение задачи, как отмечалось выше, осуществляется в целях раннего предупреждения и прогнозирования землетрясений или вулканической активности, а также изучения свойств залегающих пород в зоне размещения датчиков.

Тогда технология обработки данных должна включать в свой состав следующие модули:

1. Выбор информативного параметра.
2. Классификация текущего состояния.
3. Определение уровней и моделей взаимосвязей между данными отдельных датчиков в целях изучения свойств залегающих пород.

Ниже кратко рассматриваются основные подходы и алгоритмы, необходимые для формирования перечисленных модулей.

Выбор информативного параметра

Этот вопрос является одним из основных при формировании алгоритмов обработки данных мониторинга.

Решение описанных задач предполагает необходимость извлечения полезной информации из измеряемых сигналов. Носителями этой информации служат параметры (числовые характеристики или статистики) измеренных данных в составе матриц $\mathbf{M}_k$. Каждый $(n \times 1)$-столбец матрицы $\mathbf{M}_k$ может быть интерпретирован как реализация (результат измерения значений) случайной величины (СВ) или случайного процесса (СлПр) в каждой точке размещения датчиков. Состав списка статистических характеристик, необходимых для выбора информативного параметра, определяется типом реализации. Так, для СВ в этот состав могут войти выборочное среднее, дисперсия, медиана и другие квантили, вероятность попадания значений выборки в заданный диапазон и др. Для СлПр в этот список дополнительно могут войти максимальные значения амплитуд в заданном диапазоне частот и др.

Информативным назовем параметр, позволяющий по результатам k-го мониторинга наиболее полно охарактеризовать текущее состояние и динамику его изменения.

Следует заметить, что для каждого из m столбцов матрицы $\mathbf{M}_k$ принципиально может быть выбран свой информативный параметр. Однако однотипность датчиков и измеряемых сигналов делает возможным выбор единого информативного параметра S_k для всего массива $\mathbf{M}_k$. Стремление иметь единый для $\mathbf{M}_k$ информативный параметр объясняется также и желанием избежать необходимости решать задачу многокритериального принятия решения на заключительных этапах обработки.

При любом варианте выборки процедура выбора параметра S_k имеет общий характер. Поясним эту процедуру, предположив, что в результате мониторинга имеются два массива в виде матриц $\mathbf{M}_1$ и $\mathbf{M}_2$, которые относятся к двум различным состояниям. Для произвольных l-х ($l = 1, ..., m$) столбцов каждой из матриц $\mathbf{M}_1$ и $\mathbf{M}_2$:

- Вычисляются статистики из приведенного ранее списка; среди них получены, например, значения медиан p_1 и p_2.
- Для каждой статистики вычисляются относительные отклонения, например, $(p_2 - p_1)/p_1$.
- Относительные отклонения статистик ранжируются.
- Выбирается информативный параметр s_l l-го столбца по статистике, имеющей максимальное относительное отклонение.
- Информативный параметр S_k массива данных k-го мониторинга выбирается как статистика s_l, имеющая максимальное относительное отклонение среди l столбцов матриц $\mathbf{M}_1$ и $\mathbf{M}_2$.

Последняя позиция представляется особенно важной, поскольку обеспечивает максимальную чувствительность S_k к изменчивости состояний.

Выбранный параметр S_k достаточно прост в получении и удобен в практическом применении. Он не только отражает текущие состояния в моменты проведения мониторингов, но его последовательные значения дают возможность прогнозировать состояния между мониторингами или после их окончания. Технология прогнозирования использует принципы аппроксимации [5]. Для последовательности значений S_k в качестве аппроксимирующих выражений удобно использовать звенья динамических систем, которые описываются дифференциальными или разностными уравнениями. Решение задач аппроксимации подробно рассматривается в работах [4, 5].

Классификация состояний

Обработка данных на этом этапе заключается в сопоставлении текущих значений S_k с пороговыми значениями, априорно принятыми для угрожающей, опасной, критической и др. категорий. Решение задачи классификации может не ограничиваться параметрическим анализом. Большую помощь при решении могут оказать визуальные наблюдения за внешними проявлениями ситуаций в данном регионе (появлением трещин на земной коре, оползней, вулканических выбросов и проч.).

Оценка взаимосвязей между данными отдельных датчиков

Изучения свойств залегающих пород является одной из основных задач геофизики [1]. Методы и средства решения этих задач постоянно совершенствуются. При этом существенным представляется поиск новых подходов к определению параметров – индикаторов свойств пород в зоне их изучения. Один из таких подходов предлагается ниже. В качестве теоретической базы этот подход использует специальные разделы регрессионного и корреляционного анализа математической статистики.

Обладая сравнительной простотой, эти методы обеспечивают определение количественных оценок особенностей пород в зоне расположения датчиков.

Корреляционный и регрессионный анализы тесно связаны, имеют ряд общих параметров. Первый из методов призван оценивать степень взаимосвязи

двух (или более) СВ при различных законах (моделях) их связи. Второй метод определяет вид и параметры модели связи СВ.

В нашем случае анализу подвергаются данные, полученные от датчиков, которые здесь выступают в качестве генераторов значений (реализаций) отдельных СВ. Параметры, определяющие степень связи данных каждой пары датчиков, неизбежно становятся носителями информации о свойствах и специфике пород в зоне размещения датчиков.

Такая специфика может касаться, например, оттенков плотности пород, наличия или отсутствия трещин, пустот, жидкостных линз и проч. Датчики при этом могут замерять как естественные сигналы, так и сигналы – реакции на активное воздействие человека на земную кору. Данные о связях при расположении датчиков на поверхности или на одной глубине принципиально позволяют построить плоские (2D) модели состава пород. При размещении датчиков на разновеликих глубинах появляется возможность формирования 3D-моделей.

Формирование и градуировка шкалы плотностей пород по значениям параметров взаимосвязей может производиться на специальных стендах или полигонах с известным составом пород. При этом ясно, что повышение плотности пород будет вызывать рост уровня связи данных. В процессе установления соответствий между параметрами связи и свойствами пород должны определяться интервальные оценки для параметров взаимосвязей, что позволит принимать решения о типах пород на основе доверительных интервалов с заданным уровнем значимости [4].

Следует отметить, что значения ряда параметров связи зависят от ее модели, априорно неизвестной. Поэтому выбор параметра, инвариантного к виду модели связи представляется особенно важным при анализе свойств пород. Один из таких инвариантных параметров предлагается ниже.

Необходимость оценивать уровень связи данных, снимаемых с совокупности датчиков, возникает во многих прикладных задачах. Чаще всего в качестве параметра связи, без всяких обоснований, исследователями используется коэффициенты корреляции r [3, 4]. Однако параметр r не может служить объективной мерой уровня связи двух СВ, поскольку область его применения ограничена лишь прямолинейными моделями [4]. Это означает, что r можно использовать для оценки уровня связи двух СВ Y и X (со значениями y и x) только в случаях, когда для этих СВ справедливо уравнение прямой линии:

$$y = a + b \cdot x, \qquad (1)$$

где a и b – коэффициенты, подлежащие определению в задаче регрессии.

В случаях, когда модель связи отличается от выражения (1), применение r приводит к ошибкам, искажающим истинную картину связи двух СВ.

Для оценки этих связей можно предложить другой параметр – корреляционное отношение R [4, 5], который не имеет указанных выше ограничений и способен объективно характеризовать уровни связи при произвольных моделях, как линейных, так и нелинейных.

В отличие от r, корреляционное отношение определяется только по результатам решения задачи регрессии [4]. Сущность этой задачи для двух СВ Y и X состоит в определении регрессионного соотношения (модели связи СВ), наилучшим образом соответствующего данным наблюдения (реализациям) за этими СВ. Простейшим регрессионным соотношением служит выражение (1), причем в этой записи y и x носят названия отклика и фактора соответственно.

На рис. 2 приведены некоторые результаты решения двух задач регрессии в виде копий Mathcad-файлов – для полиномов первого порядка (1) и второго порядка вида:

$$y = a + b \cdot x + c \cdot x^2, \qquad (2)$$

где a, b и c – коэффициенты, подлежащие определению.

При решении (см. рис. 2а), для каждой модели, методом наименьших квадратов получены коэффициенты полиномов (векторы β), вычисляются остаточные суммы квадратов ε и остаточные дисперсии d, по значениям которых можно судить о качестве решения и преимуществе варианта (2).

На графиках (см. рис. 2б) точками изображены исходные реализации; линиями представлены функции (1) и (2) при найденных значениях коэффициентов соответствующих полиномов.

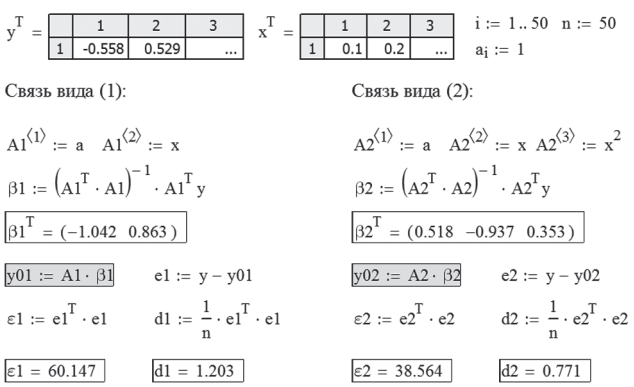
Результаты задач регрессии позволяют получить корреляционное отношение R [4]:

$$R^2 = 1 - d / (\sigma_{ув})^2, \qquad (3)$$

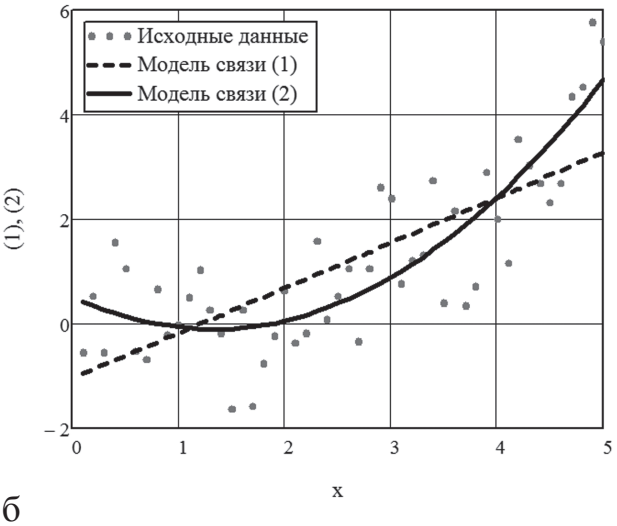
где $(\sigma_{ув})^2$ – выборочная дисперсия СВ Y.

Корреляционное отношение может служить индикатором отличия модели связи двух СВ от варианта (1), что следует из основных свойств параметров R и r:

1. В общем случае $0 \le r^2 \le R^2 \le 1$.
2. Если $r^2 = R^2 = 0$, то пара СВ некоррелирована.
3. Если $r^2 = R^2 = 1$, то пара СВ имеет линейную функциональную связь вида (1) с неслучайными a и b;
4. При $r^2 < R^2 = 1$ пара СВ связана нелинейной функциональной зависимостью (с неслучайными коэффициентами).



а



б

Рис. 2. Результаты решения двух задач регрессии:
а – определение параметров регрессии
б – графики результатов задачи регрессии

5. Если $r^2 = R^2 < 1$, то связь пары СВ линейная вида (1).
6. Если $r^2 < R^2 < 1$, то пара СВ имеет нелинейную связь (вида (2) и сложнее).

Таким образом, для задач (см. рис. 2) получены следующие результаты:

полином (1):

$R = 0.751$; $r = 0.751$; $d = 1.203$; $\varepsilon = 60.147$;

полином (2):

$R = 0.849$; $r = 0.751$; $d = 0.771$; $\varepsilon = 38.564$.

Из этих данных следует, что полином (2) представляет модель связи, более соответствующую исходным реализациям. Преимущества модели (2) хорошо видны на графиках рис. 2б и подтверждаются значениями R, d, ε.

При определении модели связи имеется возможность ее уточнения. Это уточнение можно осуществить в двух направлениях.

В качестве первого рассмотрим усложнение регрессионного соотношения. Стремление автоматизировать обработку данных мониторинга делает необходимым использование однотипных выражений в качестве регрессионных. Эти выражения выбираются на этапе предварительного анализа реализаций. Их усложнение будет приводить к увеличению значения R и снижению значений d, ε. В какой-то момент дальнейшие усложнения перестанут сопровождаться заметными изменениями параметров R, d, ε, что дает основания для выбора окончательного варианта модели. Для полиномиальной регрессии, к которой относятся выражения (1) и (2), усложнения связаны с повышениями порядка полинома.

Вторая возможность уточнить модель связи возникает при замене реализации-фактора на реализацию-отклик и наоборот. Смена ролей в паре реализаций при всех моделях, за исключением варианта (1), неизбежно вызовет изменение значений R, d, ε. Наилучший вариант модели будет соответствовать максимальному значению R и минимальным значениям d и ε. Каждый из этих параметров является критерием качества регрессии, поэтому выбор модели можно проводить по значениям одного из них, например по R.

При выборе уточненных вариантов моделей парных связей по результатам мониторинга удобно использовать $(m \times m)$-матрицу $\mathbf{P}$; здесь m – число столбцов матрицы $\mathbf{M}_k$, введенной ранее. Структура матрицы $\mathbf{P}$, подобно классической ковариационной матрице [4], в диагонали содержит выборочные дисперсии столбцов матрицы $\mathbf{M}_k$. Однако недиагональные элементы (P_{ij} и P_{ji}, $i \neq j$) матрицы $\mathbf{P}$ – значения R для $m(m-1)$ пар столбцов матрицы $\mathbf{M}_k$. Поскольку $P_{ij} = P_{ji}$, только для моделей вида (1), то, в отличие от ковариационной матрицы, матрица $\mathbf{P}$ в общем случае – несимметричная. Это ее свойство позволяет выбрать лучший вариант модели парной связи и однозначно определить, какая из реализаций в каждой паре служит фактором, а какая откликом [6]. Этот выбор осуществляется по максимальному значению в парах $P_{ij} \neq P_{ji}$.

Типовой пример матрицы $\mathbf{P}$ приведен на рис. 3.

В приведенной матрице (см. рис. 3) элементы $P_{21} > P_{12}$. Это означает, что наилучший вариант модели связи данных первого и второго датчиков

	1	2	3	4	5
1	1.32	0.24	0.32	0.68	0.13
2	0.75	2.31	0.11	0.75	0.31
3	0.21	0.15	3.42	0.13	0.63
4	0.67	0.71	0.12	3.87	0.18
5	0.35	0.76	0.12	0.21	2.12

Рис. 3. Матрица корреляционных отношений

реализуется при использовании первого столбца матрицы $\mathbf{M}_k$ в качестве отклика, а второго столбца этой матрицы – в качестве фактора.

Однако для другой пары, например P_{41} и P_{14}, можно заметить их примерное равенство: $P_{41} \approx P_{14}$. Отсюда следует, что первый и четвертый столбцы матрицы $\mathbf{M}_k$ связаны моделью вида (1), для которой назначение фактора и отклика является произвольным.

В заключение отметим, что предлагаемый подход отнюдь не исключает необходимость применения традиционных технологий при решении задач геофизики. В статье лишь делается попытка обратить внимание на свойства некоторых методов математической статистики, способных, по мнению авторов, принести пользу при решении прикладных задач.

Ключевые слова: распределенный мониторинг, математическая статистика, корреляционный анализ, регрессионный анализ, корреляционное отношение.

ЛИТЕРАТУРА

1. Кугаенко Ю.А., Титков Н.Н., Салтыков В.А. и др. Анализ подготовки трещинного толбачинского извержения 2012–2013 гг. в параметрах сейсмического режима и деформаций земной коры по данным системы комплексного мониторинга активности вулканов камчатки // Вулканология и Сейсмология. – 2015. – № 4. – С. 40-58.

2. Хмелевской В.К., Костицын В.И. Основы геофизических методов. – Пермь : Перм. ун-т, 2010. – 400 с.

3. Крамер Г. Математические методы статистики / пер. с англ. под ред. А. Н. Колмогорова. – М. : Мир, 1975. – 648 с.

4. Ивановский Р.И. Теория вероятностей и математическая статистика. Основы, прикладные аспекты с примерами и задачами в среде Mathcad. – СПб. : БХВ, 2008. – 528 с.

5. Ивановский Р.И. Статистическое моделирование. – СПб. : СПбГПУ, 2012. – 257 с.

6. Новожилов М.А. Анализ причинно-следственных связей на основе корреляционных отношений // Навигация и управление движением : материалы XVII конференции молодых ученых. – 2015. – С. 228-232.