

МАТЕМАТИЧЕСКИЕ МОДЕЛИ РАСТУЩИХ СИСТЕМ

Ю. Г. ДМИТРИЕВ, Ю. К. УСТИНОВ

Томский государственный университет, Россия

Отдел проблем информатизации ТНЦ СО РАН, Томск, Россия

e-mail: dmit70@mail.ru, ustinov@tsu.math.ru

A definition of a growing system is introduced. Corresponding mathematical models are developed for such systems. A theory of the growth processes in the accumulation systems is presented. Some examples from geology are given. Regression models are considered, and their applications to the prognostication tasks are presented.

Введение

Под растущей системой мы понимаем любую совокупность объектов одной природы, совместно участвующих в процессах зарождения, формирования, развития системы, подчиняющихся определенным, одним и тем же для всех объектов законам. Природа этих законов может быть весьма различной — полевая (гравитационная, электромагнитная, тепловая, химическая, световая и т. п.), биологическая, психологическая, экономическая, социальная, статистическая и др. Каждый из этих видов обладает разнообразными специфическими свойствами, изучение и описание которых представляет собой интересную, важную и порой очень трудную задачу.

Здесь мы рассматриваем теорию растущих систем, зарождение и развитие которых поддается статистическому, вероятностному описанию. Это в первую очередь системы, формируемые путем накопления. В геологии таковыми являются бассейны месторождений некоторого полезного ископаемого, например углеводородов, металлов, угля и т. п. В процессе охлаждения Земли в некоторых регионах ее сложились благоприятные условия для генерации, аккумуляции и накопления полезного ископаемого, обеспеченные “транспортом” сырья к месту генерации и аккумуляции. Вследствие неоднородности внутреннего строения мантии Земли, а также неровности ландшафта места генерации и аккумуляции полезного ископаемого оказались в неравноправном положении, которое можно охарактеризовать различной “приоритетностью” мест генерации и аккумуляции. Можно считать, что к настоящему времени произошло громадное число актов накопления случайных количеств вещества, которые теперь суммарно формируют запасы месторождений полезных ископаемых в бассейне и которые поддаются вероятностному описанию.

Системы подобного типа помимо геологии можно найти в биологии, социологии, экономике, политологии, демографии, астрономии и в других областях знаний.

1. Математическая модель накопления вещества в растущей системе

Рассмотрим здесь модель растущей системы, нашедшую разнообразные применения в геологии [1]. Перенумеруем все месторождения полезных ископаемых в бассейне по убыванию приоритета $1, 2, \dots, N$ (т.е. по убыванию накопленных запасов), и пусть q_k^j — весовое количество ископаемого, отложившееся в k -м месторождении в течение j -го периода, $j = \overline{1, L+1}$; Q_k^j — количество вещества, накопившееся в k -м месторождении за j периодов. Тогда, очевидно, $Q_k^j = q_k^1 + \dots + q_k^j$, $j = \overline{1, L+1}$, $Q_k = Q_k^{L+1}$ — величина запаса в k -м месторождении в настоящее время. В силу различия приоритетов мест накопления при всех $j = \overline{1, L+1}$ имеем

$$q_1^j > q_2^j > \dots > q_N^j; \quad (1)$$

$$Q_1^j > Q_2^j > \dots > Q_N^j \quad (2)$$

при всех $k = \overline{1, N}$: $Q_k^1 < Q_k^2 < \dots < Q_k^{L+1}$. В силу отмеченной выше случайной изменчивости процесса накопления мы вправе считать величины q_k^j и Q_k^j случайными. Наша задача состоит в том, чтобы найти законы их распределений.

Для описания процесса формирования Q_k^j примем за основу следующую модель накопления:

$$Q_k^{j+1} = Q_k^j + z_k^j g_k(Q_k^j), \quad j = \overline{1, L}, \quad (3)$$

где z_k^j — независимые случайные величины с математическим ожиданием $Mz_k^j = a_k > 0$ и дисперсией $Var(z_k^j) = b_k^2$, а $g_k(x)$ — некоторые функции, которые будут выбираться из физических соображений. Чтобы сохранить отношения приоритетности (1) и (2), потребуем, чтобы при $x > 0$ выполнялись условия

$$g_k(x) \geq 0, \quad g_1(x) \geq g_2(x) \geq \dots \geq g_N(x).$$

В частности, при одинаковом “механизме накопления” функции $g_k(x)$ от номера месторождения не зависят, поэтому можно принять $g_k(x) = g(x)$, где $g(x)$ — положительная, монотонно возрастающая функция. Между приращениями $q_k^{j+1} = z_k^j g_k(Q_k^j)$ при разных k и некоторых j может не выполняться требование приоритетности (1), однако оно, очевидно, выполняется в среднем.

Описанная схема накопления приводит к распределению Кептейна $\mathcal{K}(a, \sigma^2, G)$ [2, с. 129], имеющего плотность распределения

$$k(a, \sigma^2, G; x) = \frac{1}{\sigma \sqrt{2\pi} P\{\zeta \in VG\}} \exp \left\{ -\frac{(G(x) - a)^2}{2\sigma^2} \right\} \left| \frac{dG(x)}{dx} \right| I_{DG}(x). \quad (4)$$

Здесь $G(x)$ — монотонная дифференцируемая функция с областью определения DG и множеством значений VG ; ζ имеет нормальное распределение $\mathcal{N}(a, \sigma^2)$. Покажем, что если $\xi \in \mathcal{K}(a, \sigma^2, G)$, то случайная величина $\eta = G(\xi)$ имеет усеченное на множестве VG нормальное распределение $\mathcal{N}(a, \sigma^2, VG)$. В самом деле, если $y < VG$, то $P\{\eta \leq y\} = 0$. Если $y \in VG$, то

$$\begin{aligned} P\{\eta \leq y\} &= P\{G(\xi) \leq y\} = P\{\xi \in DG \cap G^{-1}(-\infty, y)\} = \\ &= \int_{DG \cap G^{-1}(-\infty, y)} \frac{1}{\sigma \sqrt{2\pi} P\{\zeta \in VG\}} \exp \left\{ -\frac{(G(x) - a)^2}{2\sigma^2} \right\} \left| \frac{dG(x)}{dx} \right| dx. \end{aligned}$$

Замена $G(x) = t$ приводит к

$$\int_{VG \cap (-\infty, y)} \frac{1}{\sigma \sqrt{2\pi} P\{\zeta \in VG\}} \exp \left\{ -\frac{(t-a)^2}{2\sigma^2} \right\} dt.$$

Отсюда для $y > VG$ получаем

$$P\{\eta \leq y\} = P\{\eta \in VG\} = \frac{1}{P\{\zeta \in VG\}} \int_{VG} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(t-a)^2}{2\sigma^2} \right\} dt = 1.$$

Таким образом, параметры a и σ^2 распределения Кептейна не имеют смысла среднего и дисперсии усеченной нормальной случайной величины $\eta = G(\xi)$.

Укажем на связь распределения Кептейна со схемой накопления (3). Пусть мы имеем накапливающуюся сумму по правилу

$$x^{j+1} = x^j + z^j g(x^j), \quad j = \overline{1, L}, \quad (5)$$

где z^j — независимые случайные величины. Будем считать для простоты, что все z^j одинаково распределены со средним a и дисперсией b^2 . Тогда

$$z^j = \frac{x^{j+1} - x^j}{g(x^j)} \quad \text{и} \quad \frac{1}{b\sqrt{L}} \sum_{j=1}^L (z^j - a) = \frac{1}{b\sqrt{L}} \sum_{j=1}^L \left(\frac{x^{j+1} - x^j}{g(x^j)} - a \right).$$

В силу центральной предельной теоремы Леви — Линдеберга [2, с. 71] левая часть здесь стремится к случайной величине $\eta \in \mathcal{N}(0, 1)$. Правая же часть легко преобразуется к виду

$$\frac{1}{b\sqrt{L}} \sum_{j=1}^L \frac{x^{j+1} - x^j}{g(x^j)} - \frac{a\sqrt{L}}{b}.$$

Если считать разности $x^{j+1} - x^j$ достаточно малыми (этого можно добиться уменьшением длительности периодов накопления вещества), то мы имеем здесь интегральную сумму. Обозначим через $G(x)$ функцию $\int_{x_0}^x \frac{du}{g(u)}$ и получаем при больших L

$$G(x^L) \approx b\sqrt{L}\eta + aL.$$

Отсюда следует, что случайная величина x^L распределена приблизительно как $\mathcal{K}(aL, b^2L, G)$.

Рассмотрим несколько примеров.

Пример 1. Пусть $g(x) = \alpha > 0$, $x > 0$. Тогда

$$G(x) = \int_0^x \frac{du}{\alpha} = \frac{x}{\alpha}. \quad (6)$$

Подставив (6) в (4), получим плотность распределения x^L :

$$f(x) \approx \frac{1}{\alpha b \sqrt{2\pi L}} \exp \left\{ -\frac{(x - \alpha L)^2}{2\alpha^2 b^2 L} \right\} \frac{1}{P\{\zeta > 0\}} I_{[0, \infty)}(x),$$

$\zeta \in \mathcal{N}(aL, b^2L)$. Это означает, что

$$x^L \in \mathcal{N}(\alpha aL, \alpha^2 b^2L, [0, \infty)).$$

Пример 2. Пусть $g(x) = \alpha x$, $x > 0$, $\alpha > 0$. Тогда $G(x) = \frac{1}{\alpha} \ln \frac{x}{x_0}$ при $x > 0$. Возьмем здесь $x_0 = 1$. Теперь случайная величина имеет распределение Кептейна с плотностью

$$f(x) \approx \frac{1}{x\alpha b\sqrt{2\pi L}} \exp \left\{ -\frac{(\ln x - \alpha aL)^2}{2\alpha^2 b^2L} \right\}, \quad x > 0.$$

Здесь $VG = (-\infty, +\infty)$. Это означает, что x^L имеет распределение, близкое к логнормальному с параметрами $\alpha aL, \alpha^2 b^2L$ и моментами

$$Mx^L \approx \exp \left\{ \frac{1}{2}\alpha^2 b^2L + \alpha aL \right\},$$

$$Varx^L \approx \exp \{ \alpha^2 b^2L + 2\alpha aL \} \{ \exp(\alpha^2 b^2L) - 1 \}.$$

Пример 3. Выше мы фиксировали $g(x)$, определяли $G(x)$ и затем записывали функцию распределения случайной величины $\xi \in \mathcal{K}$ с плотностью (4). Поставим и решим обратную задачу. Пусть задана функция распределения $H(x)$. Как следует выбрать функции $g(x)$ и $G(x)$ так, чтобы функция распределения случайной величины $\xi \in \mathcal{K}$ оказалась в точности равной H ?

Итак, пусть $\xi \in \mathcal{K}(a, \sigma^2, G) = H$ (функция G здесь еще не выбрана). Тогда по доказанному случайная величина $\eta = G(\xi) \in \mathcal{N}(a, \sigma^2, VG)$, так что

$$P\{\eta \leq y\} = P\{G(\xi) \leq y\} = P\{G(\xi) \in VG \cap (-\infty, y)\}.$$

При $y \leq VG$ пересечение $(-\infty, y) \cap VG = \emptyset$, так что $f_\eta(y) = 0$. При $y \in VG$

$$\begin{aligned} P\{\eta \leq y\} &= \int_{-\infty}^y \frac{I_{VG}(t)}{\sqrt{2\pi}\sigma P\{\zeta \in VG\}} \exp \left\{ -\frac{(t-a)^2}{2\sigma^2} \right\} dt = \\ &= \frac{N(a, \sigma^2, y) - N(a, \sigma^2, \min VG)}{P\{\zeta \in VG\}} = P\{-\infty \leq \xi \leq G^{-1}(y)\} = H(G^{-1}(y)). \end{aligned}$$

Переобозначим здесь $G^{-1}(y) = t$, $y = G(t)$, так что

$$N(a, \sigma^2, G(t)) = N(a, \sigma^2, \min VG) + P\{\zeta \in VG\}H(t) = \Phi \left(\frac{G(t) - a}{\sigma} \right).$$

Отсюда

$$G(t) = a + \sigma \Phi^{-1}[N(a, \sigma^2, \min VG) + P\{\zeta \in VG\}H(t)].$$

Вернемся к модели накопления (3), которая имеет форму (5). Согласно полученным выше результатам, накопленный запас Q_k в k -м месторождении распределен приблизительно по Кептейну с параметрами $La_k, Lb_k^2, G_k(x)$, где

$$a_k = Mz_k, \quad b_k^2 = Dz_k, \quad G_k(x) = \int_{x_0}^x \frac{du}{g_k(u)}.$$

Данную совокупность месторождений с накопленными за длительный период времени запасами Q_1, \dots, Q_N назовем *природной совокупностью*. Выбирая различным образом функции g_k , мы можем строить модели природных совокупностей с желательным распределением запаса в k -м месторождении.

Рассмотрим вопрос о выборе функций $g_k(x)$. Если мы имеем дело с месторождением осадочного типа, то количество оседающего вещества в каждом периоде не зависит от количества накопленного. Такая ситуация описана в примере 1. В результате оказывается, что запас Q_k распределен приблизительно нормально с известными по примеру 1 параметрами. Если месторождение формировалось по закону роста кристаллов, то количество откладывающегося вещества в каждом периоде пропорционально накопленному. Эта ситуация описана в примере 2, откуда следует, что запас Q_k распределен приблизительно логнормально с известными по примеру 2 параметрами. Другие геологические механизмы отложения полезного ископаемого описываются соответствующими функциями $g(x)$.

2. Статистическая интерпретация природной совокупности

Следующий вопрос, который здесь возникает, состоит в выяснении распределения генеральной совокупности, выборку из которой образуют запасы данной природной совокупности. Проблема заключается в неопределенности относительно того, в какой мере можно считать запасы месторождений природной совокупности независимыми наблюдениями из некоторой генеральной совокупности? Рассмотрим процесс накопления вещества в природной совокупности как процесс формирования многомерного вектора $\mathbf{x} = (x_1, \dots, x_N)$. Будем считать, что прибавки вещества формируются согласно векторной функции $\mathbf{g}(\mathbf{x}) = (g_1(x_1), \dots, g_N(x_N))$ с коэффициентом $\mathbf{z} = (z_1, \dots, z_N)$, имеющим

$$M\mathbf{z} = \mathbf{a} = (a_1, \dots, a_N) \quad \text{и} \quad \text{cov}(\mathbf{z}) = (b_{ij})_{i,j=\overline{1,N}} = B.$$

Схему (3) перепишем в следующем виде: $\mathbf{x}^{j+1} = \mathbf{x}^j + \mathbf{z}^j \mathbf{g}(\mathbf{x}^j)$, $j = \overline{1, L}$. Умножение векторов \mathbf{z}^j и $\mathbf{g}(\mathbf{x}^j)$ здесь (и деление векторов ниже) принимается покоординатное:

$$\mathbf{z}^j \cdot \mathbf{g}(\mathbf{x}^j) = (z_1^j \cdot g_1(x_1^j), z_2^j \cdot g_2(x_2^j), \dots, z_N^j \cdot g_N(x_N^j)).$$

Будем считать векторы \mathbf{z}^j независимыми и одинаково распределенными. Тогда, действуя так же, как и в одномерном случае, по многомерной теореме Леви — Линдберга [2, с. 78] получим

$$\frac{1}{\sqrt{L|B|}} \left(\sum_{j=1}^L \frac{\mathbf{x}^{j+1} - \mathbf{x}^j}{\mathbf{g}(\mathbf{x}^j)} - \mathbf{a}L \right) = \frac{1}{\sqrt{L|B|}} \sum_{j=1}^L (\mathbf{z}^j - \mathbf{a}) \rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

где $L \rightarrow \infty$; \mathbf{I} — единичная матрица. Обозначим

$$G_k(x_k) = \int_{x_k^0}^{x_k} \frac{du}{g_k(u)}, \quad \mathbf{G}(\mathbf{x}) = (G_1(x_1), G_2(x_2), \dots, G_N(x_N)).$$

В этих обозначениях мы имеем

$$\frac{\mathbf{G}(\mathbf{x}^{L+1}) - \mathbf{a}L}{\sqrt{L|B|}} = \mathbf{G}^*(\mathbf{x}^{L+1}) \in \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Отсюда следует, что координаты вектора $\mathbf{G}^*(\mathbf{x}^{L+1}) = \left(G_1^*(x_1), G_2^*(x_2), \dots, G_N^*(x_N) \right)$ можно считать независимыми и $\mathcal{N}(0, 1)$ — распределенными. Иначе говоря, координаты именно этого вектора можно рассматривать как выборку из $\mathcal{N}(0, 1)$ после подстановки вместо \mathbf{x}^{L+1} наблюдений

$$\left(x_1^{*L+1}, x_2^{*L+1}, \dots, x_N^{*L+1} \right).$$

Таким образом, для k -го месторождения имеем

$$G_k(Q_k) \approx b_k \sqrt{L} \eta_k + a_k L. \quad (7)$$

Пользуясь тем, что η_k имеют стандартную нормальную функцию распределения $\Phi(t)$, данное соотношение запишем в другом виде:

$$\Phi \left(\frac{G_k(Q_k) - a_k L}{b_k \sqrt{L}} \right) \approx \Phi(\eta_k) = u_k, \quad k = \overline{1, N},$$

где u_1, \dots, u_k — равномерно распределенные в $(0, 1)$ случайные величины.

Каждой реализации Q_1, \dots, Q_N соответствуют $G_1(Q_1), \dots, G_N(Q_N)$ и выборочная реализация u_1, \dots, u_N . В силу приоритетности мест накопления мы имеем упорядоченный набор $Q_1 > Q_2 > \dots > Q_N$, чего нельзя сказать о u_1, \dots, u_N , кроме того, они на практике неизвестны, так как неизвестны a_k и b_k . По условиям формирования месторождений в природной совокупности все они являются продуктом одного “механизма накопления” и поэтому естественно описать функцию G от накопленных запасов месторождений одним распределением вероятностей. Так как в силу (7) все $G_k(Q_k)$ имеют приблизительно нормальное распределение (хотя и с разными параметрами), общее распределение естественно искать среди нормальных $[(G(Q) - \mu)/\sigma] \approx \mathcal{N}(0, 1)$, где μ и σ — подходящим образом выбранные параметры. Исходя из того принципа, что каждая случайная величина, стремясь проявиться в эксперименте наилучшим образом, “показывает” свое распределение в квантилях X_k уровней

$$p_k = \frac{2N - 2k + 1}{2N}, \quad k = \overline{1, N},$$

параметры μ и σ найдем как значения, доставляющие минимум сумме:

$$\sum_{k=1}^N \left(\Phi \left(\frac{G_k(Q_k) - \mu}{\sigma} \right) - \frac{2N - 2k + 1}{2N} \right)^2.$$

Таким образом, аппроксимирующим общим распределением природной совокупности является $\mathcal{N}(\mu, \sigma)$, а значения $G_k(Q_k)$ есть оценки квантилей $\mu + \sigma X_k$ уровней p_k .

3. Регрессионные модели и прогнозирование в природной совокупности

Статистическая интерпретация природной совокупности и формула (7) позволяют рассматривать регрессионные модели и решать задачу прогнозирования (оценки) величин запасов ее элементов. В схеме накопления примеров 1 и 2 для k -го месторождения имеем соответственно $Q_k \approx \alpha_k b_k \sqrt{L} \eta_k + \alpha_k a_k L$ и $\ln Q_k \approx \alpha_k b_k \sqrt{L} \eta_k + \alpha_k a_k L - \ln x_0$. Обобщая эти зависимости, можем записать $G(Q_k) \approx v_k \eta_k + \mu_k$, $k = \overline{1, N}$, где функция G от индекса k (номера месторождения) не зависит, а v_k, μ_k — обобщенные параметры. В случае $g_k = g$ также отсутствует зависимость от номера месторождения у функции G .

Расположим величины запасов Q_k по величине в порядке убывания и поставим им в соответствие $Y_k = G(Q_k)$. Предположим, что для всех месторождений природной совокупности средние квадратические отклонения величин Y_k одинаковы, т. е. $v_k = v$, а средние значения $\mu_k = \mu + \sigma X_k$, т. е. пропорциональны квантилям X_k уровней p_k . Тогда уравнение

$$Y_k = \mu + \sigma X_k + \varepsilon_k, \quad k = \overline{1, N}, \quad (8)$$

где μ и σ — параметры, подлежащие оцениванию; $\varepsilon_k = v \eta_k$ — независимые гауссовские величины с $M\varepsilon_k = 0$ и дисперсией $D\varepsilon_k = v^2$, задает классическую регрессионную модель. Как известно [3], в силу теоремы Гаусса — Маркова среди всех линейных и несмещенных оценок наилучшими являются оценки, полученные методом наименьших квадратов (МНК-оценки). В нашем случае гауссовских ошибок для оценивания μ, σ и v^2 можно воспользоваться методом максимального правдоподобия.

Дальнейшие обобщения регрессионной модели (8) связаны с предположениями относительно изменения средних квадратических отклонений v_k . Например, $v_k = c_k v$, где коэффициенты пропорциональности c_k можно брать в виде $c_k = p_k$, считая, что чем больше величина накопленного запаса, тем больше и дисперсия; или $c_k = p_k$ при $k < N/2$, $c_k = 1 - p_k$ при $k > N/2$, тем самым полагая, что и большим, и малым по величине запасам присущи большие дисперсии. В этих случаях получаем линейные регрессионные модели с гетероскедастичностью и для оценки параметров регрессии можно применить взвешенный метод наименьших квадратов, сводящийся к минимизации по μ и σ суммы:

$$\sum_{k=1}^N (c_k^{-1} (Y_k - \mu - \sigma X_k))^2.$$

Рассмотрим задачу прогнозирования. Пусть известно, что природная совокупность состоит из N элементов (месторождений), однако исследователю известна лишь часть природной совокупности, состоящая из $n < N$ упорядоченных элементов $Q_{(1)} > \dots > Q_{(n)}$ с неизвестными номерами, отвечающих тому или иному элементу совокупности. Нужно по имеющимся данным спрогнозировать (оценить) величины запасов оставшихся $N - n$ элементов природной совокупности, номера которых также неизвестны.

Данную задачу будем решать следующим образом. Вычислим по исходным данным упорядоченную последовательность $Y_{(1)} > \dots > Y_{(n)}$ и найдем для нее наилучшее соответствие $X_{(1)} > \dots > X_{(n)}$ из имеющихся квантилей $X_1 > \dots > X_N$. Наилучшим будем считать такое соответствие, которое доставляет минимум сумме

$$\sum_{j=1}^n (c_{(j)}^{-1} (Y_{(j)} - \mu - \sigma X_{(j)}))^2$$

по всевозможным переборам соответствий из N по n с учетом порядка, а также μ и σ . Если такое соответствие найдено, то становятся известными номера прогнозируемых элементов совокупности и соответствующие им квантили. Подставляя эти квантили в оцененное уравнение регрессии, найдем прогнозные значения неизвестных элементов совокупности. В частности, такой подход позволяет оценивать накопленные запасы полезных ископаемых еще не обнаруженных месторождений по известным величинам запасов уже разведанных месторождений.

Список литературы

- [1] АЛЕКСЕЕВ Ф.Н., РОСТОВЦЕВ В.Н. Теория образования месторождений полезных ископаемых и практика ее применения. Томск: STT, 2004. 315 с.
- [2] СПРАВОЧНИК по теории вероятностей и математической статистике / В.С. Королюк, Н.И. Портенко, А.В. Скороход, А.Ф. Турбин. М.: Наука, 1985.
- [3] МАГНУС Я.Р., КАТЫШЕВ П.К., ПЕРЕСЕЦКИЙ А.А. Эконометрика. Начальный курс: Учебник. 4-е изд. М.: Дело, 2000. 400 с.

Поступила в редакцию 23 августа 2007 г.