УДК 556.5.114 (075.8)

О СТАТИСТИЧЕСКИХ МЕТОДАХ В ПРЕДСКАЗАНИЯХ ПРИРОДНЫХ ПРОЦЕССОВ В РЕЧНОЙ ЭКОСИСТЕМЕ

Доктор техн. наук М.Ж. Бурлибаев

При современном состоянии деградации речных экосистем важным становится восстановление речного комплекса до уровня наблюдавшихся при естественном гидрологическом режиме водотоков. Для этих целей необходимо четко прогнозировать перспективные природные процессы, которые могут происходить в речной экосистеме под влиянием антропогенных процессов. Поэтому данная статья посвящается статистическим методам предсказания природных процессов в речной экосистеме.

Учет структурных особенностей при диагнозе характеристик в долгосрочном прогнозе природных процессов в речной экосистеме, возможен при следующих предпосылках: во-первых, чтобы статистические характеристики рядов были устойчивы; во-вторых, необходимо наличие определенной регулярности в проявлении цикличности. Первое условие в природных процессах выполняется весьма приближенно, что касается второго, то оно накладывает следующие ограничения – наличие в квазипериодических колебаниях закономерного чередования высоких и низких значений, а в квазислучайных (стохастических процессах) – хотя бы группировки высоких и низких значений, отличной от случайной. Чтобы выявить наличие случайности (неслучайности) колебаний, необходимо использовать критерии систематического ряда, например, критерии экстремумов, повышений или понижений ряда (критерии Б.П. Вайнберга, М.А. Омшанского, О.А. Дроздова и др.) [1, 2, 3]. Цель использования таких критериев – выявить свойства в колебаниях анализируемых рядов, отличных от колебаний, имеющих место в случайных бессвязных рядах. Отметим, что в последних (рядах) числа повышений и понижений приблизительно равны, а число экстремумов распределено асимптотически нормально со средним $m_э \approx 2n/3$ и дисперсией $D_э \approx (16n-29)/90$. Для проверки гипотезы случайности ряда x_i достаточно рассчитать фактическое нормированное число экстремумов:

$$t_э = n^*_э - m_э / \sqrt{D_э}, \quad (1)$$

и сравнить его с нормированным значением t_α нормального закона распределения при заданном уровне значимости α. При $t_э > t_\alpha$, гипотеза о случайности ряда будет, по-видимому, неверна. Сущность критерия проверки случайности по числу повышений или понижений ряда состоит в следующем. Пусть имеется выборка x_1, x_2, ..., x_n; к повышениям (+) в ряду относится ситуация, когда $x_{i-1} < x_i$, к понижениям (–) – когда $x_{i-1} > x_i$. Общее число ситуаций (+) + (–) = N распределено асимптотически нормально с математическим ожиданием:

$$m_+ = m_- = n/2. \quad (2)$$

и дисперсией

$$D_+ = D_- = n + 1/12. \quad (3)$$

Зная m и D, можно рассчитать нормированные значения повышений или понижений:

$$t^*_+ = \frac{n^*_+ - m_+}{\sqrt{D_+}}; \qquad t^*_- = \frac{n^*_- - m_-}{\sqrt{D_-}}, \quad (4)$$

где n^*_+, n^*_- – соответственно, число повышений (понижений) в ряду. Затем, проводится сравнение t^*_+ и t^*_- со значениями нормированных ординат таблицы закона нормального распределения. Если t^*_+ и t^*_- по таблице окажется меньше уровня значимости, то гипотеза о случайности исследуемого ряда отвергается и принимается гипотеза об устойчивости тенденции к повышению или понижению. Критерий Б.П. Вайнберга позволяет обнаружить изменение уровня ряда, а критерий М.А. Омшанского – оценить длительность разных циклов. Критерий случайности О.А. Дроздова базируется на вычислении разностей – $d_1 = x_2 - x_1$, $d_2 = x_3 - x_2$,…, $d_{n-1} = x_n - x_{n-1}$, между членами исходной последовательности. Суммирование разностей d_i приводит к выражению

$$D_k = \sum_{i=1}^{k} d_i = x_{k+1} - x_1, \quad (5)$$

т. е. при суммировании разностей вновь получается разностный ряд, но с интервалом в k – членов. Относя эти накопления разности к среднему

квадратическому отклонению разностного ряда σ_{Dk}, получаем искомый критерий случайности C_k по сравнению с естественной изменчивостью:

$$C_k = D_k / \sigma_D. \quad (6)$$

При наличии систематических тенденций эволюций уровня, ряд – C_k (k = 1, 2, 3,..., N), при одном из значений k, наконец, выйдет за пределы нескольких единиц, что будет характеризовать надежность установления тенденции. Мощность критерия C_k возрастает при вычислении разностей не между отдельными членами ряда, а между исследуемыми n-летиями

$$d_k^{(n)} = \frac{1}{n} \cdot \left(\sum_{i=k+1}^{k+n} x_i - \sum_{i=k}^{k+n+1} x_i \right). \quad (7)$$

Суммирование $d_k^{(n)}$ дает величины:

$$D_k^{(n)} = \frac{1}{n} \cdot \left(\sum_{i=k+1}^{k+n} x_i - \sum_{i=1}^{n} x_i \right). \quad (8)$$

Для случайного бессвязного ряда дисперсия определяется как:

$$\sigma_D^{2(n)} = 2 \cdot \sigma_x^2 / n, \quad (9)$$

откуда

$$C_k^n = D_k^{(n)} / \sigma_D^{(n)}. \quad (10)$$

Сравнивая ряд $D_k^{(n)}$ с $\sigma_D^{(n)}$, можно более точно оценить существенность отличия изменений в ряду x_i от колебаний в случайном бессвязном ряду. Математически постановку задачи статистического прогноза можно сформулировать следующим образом. Пусть $x(t)$ – стационарный процесс, наблюдавшийся до момента t_0. После t_0 сведений о значениях процесса нет. Требуется предсказать $\hat{x}(t_0 + \Theta)$ – значение процесса в момент $(t_0 + \Theta)$, причем, – с наилучшей точностью. Истинное значение процесса $x(t_0 + \Theta)$, как правило, не совпадает с предсказанным $\hat{x}(t_0 + \Theta)$. Их разность

$$e(t + \Theta) = x(t_0 + \Theta) - \hat{x}(t_0 + \Theta), \quad (11)$$

представляет ошибку прогноза на время Θ, произведенного в момент t_0. Располагая прошлыми и текущими значениями прогноза (предысторией), можно получить характеристику процесса, определяющую связь между его значениями, разделенными временным промежутком Θ, а именно – корреляционную функцию. В гидрометеорологических исследованиях понятие статистическо-

го прогноза обычно связывается с задачей экстраполяции (интерполяции) и сглаживания случайного процесса. По известной реализации:

$$z(t) = x(t) + y(t), \tag{12}$$

в которой $x(t)$ – детерминированная составляющая; $y(t)$ – случайная составляющая; в случае $y(t) = 0$ (процесс без ошибок) прогноз сводится к чистой экстраполяции. В случае наличия ошибок $y(t)$, прежде чем определить истинное значение реализации $x(t)$ в некоторый момент $t+\Theta$, необходимо отделить его от ошибки наблюдения. Это задача о сглаживании (фильтрации) случайного процесса. Задача об экстраполяции тесно связана со сглаживанием, так как реализация случайного процесса включает в себя ошибки измерения. При этом задача экстраполяции сглаживанием состоит в том, чтобы по имеющейся реализации (12) на промежутке $t_0 + \Theta$ дать прогноз реализации $x(t)$ в момент $t + \Theta$, при $\Theta > 0$. При $\Theta < 0$ – имеет место задача интерполяции со сглаживанием. В математической постановке задачи предполагается, что математические ожидания процессов – $m_x(t),\ m_y(t)$, их автокорреляционные и взаимные корреляционные функции заданы. При этом обычно считают $m_x(t)$ и $m_y(t)$ равными 0. В противном случае, вместо $x(t)$ и $y(t)$ рассматриваются их центрированные случайные функции (аномалии) – $\Delta x(t)$, $\Delta y(t)$. Математическое решение задачи статистического прогноза сводится к получению наилучшего результата по всему множеству реализаций, т. е. к нахождению такого оператора L, который в применении к множеству реализаций $z(t)$ давал бы наилучшее, в некотором смысле, значение реализации $x(t_0 + \Theta)$

$$x(t_0 + \Theta) = L(x(t) + y(t)). \tag{13}$$

Для оценки качества прогнозирования вводится критерий качества прогнозирования, так или иначе связанный с ошибкой прогноза, – средним квадратом ошибки:

$$\bar{e}^2(t_0 + \Theta) = M\left(\left(x(t_0 + \Theta) - \hat{x}(t_0 + \Theta)\right)^2\right) \Rightarrow \min. \tag{14}$$

Чтобы вычислить предсказанное значение, нужно уметь выбрать правило вычисления ожидаемой оценки $\hat{x}(t_0 + \Theta)$ – алгоритм прогноза. Алгоритм предсказания $\hat{x}$ должен связать с ним предысторию процесса и его вероятностные характеристики. С другой стороны, качество алгоритма определится дисперсией ошибки прогноза.

Рассмотрим простейшие алгоритмы прогноза.

Вероятностное прогнозирование значений случайного процесса

Оценка значения в реализации случайного процесса (в силу случайности физических явлений) в будущем (в момент времени – t) не может быть вычислена по точной формуле, но может быть описана в вероятностном виде. В случае стационарного и эргодического процессов, $P(x, t)$ не зависит от времени и может быть определена по единственной реализации $x(t)$ как:

$$P(x) = P(x(t) \le \xi) = \frac{\lim}{T \to \infty}\left(\frac{T(x(t) \le \xi)}{T}\right), \qquad (15)$$

где $T(x(t) \le \xi)$ – общее время, в течение которого реализация $x(t)$ находится не выше уровня ξ. В этом случае, значение $x(t)$ в произвольный момент времени не превышает данного значения ξ. При $x \to -\infty$, интегральная функция распределения $P(x) = P(x, t)$ стремится к 0, при $x \to \infty$ – стремится к 1. По характеру изменения функции распределения от 0 до 1 различаются случайные процессы с разной вероятностной структурой. Можно сказать, что плотность вероятности определяет скорость изменения функции распределения, поскольку вероятность различных событий можно находить интегрированием плотности вероятности на кривой:

$$P(x_1 \le x(t) \le x_2) = \int_{x_1}^{x_2} P(x) \cdot dx = P(x_2) - P(x_1). \qquad (16)$$

Функция распределения существует как для непрерывных, так и для прерывных случайных величин и является универсальной характеристикой случайных величин, так как плотность характеризует их с вероятностной точки зрения. Зная функцию распределения случайной величины, можно найти вероятность ее попадания на заданный участок, которая равна приращению функции распределения на этом участке. При $x_1 \to 0$, получим x_2,

$$P(0 \le x(t) \le x_2) = \int_{-\infty}^{x_2} P(x) \cdot dx = P(x_2), \qquad (17)$$

т. е. площадь под графиком плотности вероятности левее точки x_2 равна значению дифференциальной функции распределения в точке x_2. Таким образом, прогнозирование вероятности того или иного элемента может быть осуществлено при знании или прогнозировании функции распределения. Задача прогнозирования при использовании вероятностных моделей заключается в определении по кривой распределения вероятностей величины параметра x, такого, когда вероятность $P(x)$ равна заданному значению P. Следует пом-

нить, что точность прогноза, с вероятностной точки зрения, в этом случае, будет зависеть от точности прогноза функции распределения.

Прогноз по последнему значению

Прогнозирование по последнему значению реализации (инерционный прогноз), заключается в том, что в качестве предсказанного значения $\hat{x}(t_0 + \Theta)$ принимается значение $x(t_0)$:

$$\hat{x}(t_0 + \Theta) = x(t_0). \quad (18)$$

Предсказанное значение здесь не зависит от предыстории прогноза (предыстория представлена лишь одной точкой – последним значением $x(t_0)$), а вероятностные характеристики не учитываются совсем. Алгоритм прогноза заключается в умножении значения последнего наблюдения $x(t_0)$ на 1, т. е. не требует выполнения никаких вычислительных операций. Прогноз, таким образом, можно выполнить, ничего не зная о процессе, не производя никаких вычислений. Однако, точность прогноза очень низкая. Возможная ошибка прогноза по алгоритму (11) здесь определяется как:

$$e(t_0 + \Theta) = x(t_0 + \Theta) - x(t_0), \quad (19)$$

а ее средний квадрат $\bar{e}^2$, если $m_x = 0$,

$$\bar{e}^2(\Theta) = M((x(t_0 + \Theta) - x(t))^2) = \sigma_x^2 - 2 \cdot r_x(\Theta) + \sigma_x^2 = 2 \cdot (\sigma_x^2 - r_x(\Theta)), \quad (20)$$

Средний квадрат ошибки прогноза растет от 0, при $\Theta = 0$, когда $r_x(0) = \sigma_x^2$, до $2\sigma_x^2$, при $\Theta = \infty$, когда $r_x(\infty) = 0$. Об истинном качестве этого способа прогноза можно говорить после сравнения полученной ошибки с ошибками других алгоритмов и способов прогноза. Простота этого способа обеспечила ему широкое распространение.

Прогноз по математическому ожиданию

Прогнозирование по математическому ожиданию заключается в том, что в качестве предсказанного значения $\hat{x}(t_0 + \Theta)$ принимается математическое ожидание m_x. Как и в предыдущем способе, предсказанное значение здесь не зависит от времени прогноза Θ. Различие заключается в том, что, хотя не требуется никакой информации о предыстории, необходимы сведения о свойствах процесса – о его математическом ожидании. Алгоритм прогноза не требует никаких вычислительных операций. Ошибка прогноза вычисляется по зависимости:

$$e(\Theta) = x(t + \Theta) - m_x, \quad (21)$$

и представляет собой отклонение процесса от среднего в момент $t_0 + \Theta$.

Средний квадрат ошибки не зависит от времени прогноза и равен дисперсии прогноза:

$$\bar{e}^2(\Theta) = M((x(t_0 + \Theta) - m_x))^2) = \sigma_x^2 . \qquad (22)$$

При малой заблаговременности Θ прогноз по последнему значению явно предпочтителен, однако, после получения некоторой критической величины заблаговременности прогноза Θ^*, когда $\bar{e}^2(\Theta^*) = \sigma_x^2$, метод прогноза по математическому ожиданию дает большую точность. Наконец, при $\Theta \to \infty$ квадрат ошибки прогноза по математическому ожиданию (норме) вдвое меньше, чем по последнему отсчету. Так, предсказывая расход воды в реке на несколько дней, мы, руководствуясь инерцией, ориентируемся на ее текущее состояние, совершенно игнорируя средние многолетние величины. Наоборот, пытаясь предвидеть летом весеннее половодье, мы, напротив, прежде всего, ориентируемся на «норму» половодья.

Статистический прогноз по одной точке

Стационарный эргодический процесс может быть как ансамблем реализаций, так и одной реализацией неограниченной длительности. Сечения ансамбля представляют собой случайные величины, функция распределения которых отождествляется с одномерной функцией распределения процесса. Обозначим случайную величину $x(t_0)$ как сечение процесса в момент t_0 – через x, а сечение $x(t_0 + \Theta)$ – через y и будем рассматривать систему двух случайных величин x, y – последнего значения предыстории и предсказанного значения. Компоненты системы x и y подчинены одномерным нормальным законам – $N_1(m_x, \sigma_x)$, $N_2(m_y, \sigma_y)$. Алгоритм прогноза в рассматриваемом способе формулируется так, что в качестве предсказанного значения $\hat{x}(t_0 + \Theta)$ выступает условное математическое ожидание $m_{y/x}$ величины y, при условии, что $x = x(t_0)$

$$\hat{x}(t_0 + \Theta) = m_{y/x}, \qquad (23)$$

где аналогично уравнению регрессии:

$$m_{y/x} = m_y + r\frac{\sigma_y}{\sigma_x} \cdot (x - m_x); \qquad (24)$$

$$\sigma_{y/x} = \sigma_y \cdot \sqrt{1-r^2}. \quad (25)$$

Известно, что закон, постулированный при условии, что первая компонента *x* приняла определенное значение, называется условным законом распределения и имеет вид:

$$f_{(y/x)} = \frac{f(x,y)}{f(x)} = \frac{1}{\sigma_y \cdot \sqrt{2\cdot\pi} \cdot \sqrt{1-r^2}} \cdot \exp\left(-\frac{1}{2(1-r^2)}\right) \cdot \left(\frac{y-m_y}{\sigma_y} - r \cdot \frac{x-m_x}{\sigma_x}\right) \quad (26)$$

или

$$f_{(y/x)} = \frac{1}{\sigma_y \cdot \sqrt{2\cdot\pi} \cdot \sqrt{1-r^2}} \cdot \exp\left(-\frac{1}{2(1-r^2)}\right) \cdot \left(y - m_y - r \cdot \frac{\sigma_y}{\sigma_x} \cdot (x-m_x)\right). \quad (27)$$

С учетом (26) и (27), получим плотность нормально распределенной условной случайной величины:

$$f_{(y/x)} = \frac{1}{\sqrt{2\cdot\pi} \cdot \sigma_{y/x}} \cdot \exp\left(-\frac{y-m_{y/x}}{2\cdot\sigma_{y/x}^2}\right). \quad (28)$$

Из (28) следует, что, при изменении одной из компонент, вид закона распределения второй компоненты не меняется, а меняется лишь его параметр $m_{y/x}$ – условное математическое ожидание (24). Условная дисперсия (25) от значения *x* также не зависит. Зависимость $m_{y/x}$ от *x* линейна и называется регрессией *y* на *x*. Ошибка прогноза по соотношению (24) определяется как

$$e(t_0+\Theta) = x(t_0+\Theta) - \hat{x}(t_0+\Theta) = y - m_{y/x} \quad (29)$$

и представляет отклонение случайной величины *y* от своего условного математического ожидания, а средний квадрат ошибки – $\bar{e}^2(\Theta) = M(y-m_{y/x})^2$ равен условной дисперсии $\sigma_{y/x}^2$. Учитывая (25) и (26), установим, что:

$$\hat{x}(t_0+\Theta) = m_{x/y} = m_y + r_{x/y} \cdot \frac{\sigma_y}{\sigma_x} \cdot (x-m_x) \quad (30)$$

и

$$\bar{e}^2(\Theta) = \sigma_{y/x} = \sigma_y^2 \cdot (1-r^2). \quad (31)$$

Поскольку процесс *x*(*t*) стационарен, математические ожидания и дисперсии сечений одинаковы: $\sigma_y = \sigma_x = \sigma$; $m_y = m_x = m$. Коэффициент

корреляции r_{xy} равен значению нормированной корреляционной (автокорреляционной) функции

$$r_{xy} = r(\Theta). \tag{33}$$

Теперь алгоритм прогноза прост:

$$\hat{x}(t_0 + \Theta) = m + r(\Theta) \cdot (x(t_0) - m), \tag{34}$$

а средний квадрат ошибки оказывается зависящим от Θ:

$$e^{-2} = \sigma^2 \cdot (1 - r(\Theta)). \tag{35}$$

Алгоритм предполагает знание отклонения процесса от среднего в момент t_0, т. е. одной точки предыстории, знание нормированной корреляционной функции $r(\Theta)$ и математического ожидания m.

Динамико-стохастический метод сверхдолгосрочного прогноза

Математический аппарат прогнозирования динамико-стохастическим методом содержит в своей основе предположение о стационарности прогнозируемых процессов во времени и наличии в процессе внутрирядных связей (даже при сдвиге степени τ>1). Задача линейного экстраполирования (прогноза) стационарной случайной последовательности, удовлетворяющей условиям: m_x = const, $r_x = r(\tau) \neq 0$, заключается в необходимости подбора таких действительных коэффициентов κ_τ, при заданных $m > 0$ и $\Theta > 0$, при которых линейная комбинация:

$$\hat{x}_0(t_0 + \Theta) = \bar{x} + k_1 \cdot \Delta x_{t-1} + k_2 \cdot \Delta x_{t-2} + \ldots + k_m \cdot \Delta x_{t-m} = \bar{x} + \sum_{\tau=1}^{m} k_\tau \cdot \Delta x_{t-\tau}, \tag{36}$$

является наиболее точным приближением к случайной величине $\hat{x}(t)$. В соотношении (36): $\bar{x}_0(t_0 + \Theta)$ – прогнозируемое на момент $t = t_0 + \Theta$ значение исследуемой величины $x(t)$; $\bar{x}$ – среднее значение $x(t)$; Δx_{t-1}, Δx_{t-2}, ..., Δx_{t-m} – последовательность предшествующих моменту t значений величины $x(t)$ в отклонениях от $\bar{x}$; k_1, k_2, ..., k_m – коэффициенты обратной связи, определяемые путем решения системы уравнений:

$$\left.\begin{array}{l} k_1 + k_2 \cdot r_2 + \ldots + k_m \cdot r_{m-1} = r_1, \\ k_1 \cdot r_1 + k_2 + \ldots + k_m \cdot r_{m-2} = r_2, \\ \ldots \quad \ldots \quad \ldots \\ k_1 \cdot r_{m-1} + k_2 \cdot r_{m-2} + \ldots + k_m = r_m \end{array}\right\}, \tag{37}$$

где r_1, r_2, ..., r_m – последовательность значений корреляционной функции $r(\tau)$; m – оптимальный период обратной связи.

Для определения периода обратной связи $m_{опт}$ рекомендуется выполнять проверочное прогнозирование, при значениях m, последовательно увеличивающихся от 1 до 30 лет. В качестве оптимального выбирается то значение, при котором ошибка прогноза $\bar{e}^{-2}$ становится минимальной. Оценка точности прогнозов, как правило, производится по последовательности эмпирических коэффициентов связи между фактическими $\Delta x_{iф}$ и прогностическими Δx_{ip} аномалиями ряда. При этом прогностическое значение $\Delta x(t + \theta)$ отыскивается по уравнению авторегрессии вида:

$$\Delta x(t + \Theta) = \sum_{k=0}^{m} \alpha_k \cdot \Delta x(t - k). \qquad (38)$$

Коэффициенты α_k для каждого заданного значения θ, определяются, исходя из условия минимума ошибки экстраполяции, при решении системы уравнений:

$$r_{\Delta x}(\Theta + j) = \sum_{k=1}^{m} \alpha_k \cdot r_{\Delta x}(k - j); \text{ при } j = 1, 2, ..., m, \qquad (39)$$

где $r_{\Delta x}(\tau)$ – корреляционная функция отклонений.

Число слагаемых m в сумме $\sum_{k=1}^{m}$ следует выбирать таким, чтобы корреляционные моменты $r_{\Delta x}(k - j)$ определялись по данным наблюдений в m – точках с требуемой надежностью. На рис. приведены результаты прогноза годового стока реки с заблаговременностью 1 год.

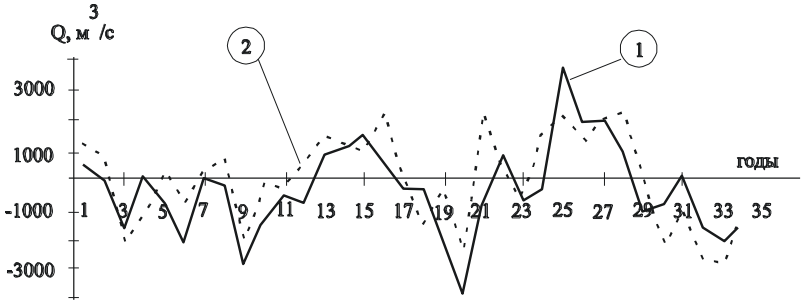

Рис. Результаты прогноза годового стока реки с заблаговременностью один год: 1 – фактический сток, 2 – прогнозные величины стока.

СПИСОК ЛИТЕРАТУРЫ

1. Бурлибаев М.Ж., Волчек А.А., Шведовский П.В. Проблемы оптимизации природопользования и природообустройства в математических моделях и методах. – Алматы: Изд-во: «Каганат», 2003. – 532 с.
2. Бурлибаев М.Ж., Нурмаганбетов Д.Ш., Волчек А.А. Теоретические и прикладные основы проблем планирования и управления природопользованием и охраной природы. – Алматы, Изд-во: «Каганат», 2007. – 360 с.
3. Бурлибаев М.Ж. Теоретические основы устойчивости экосистем трансзональных рек Казахстана. – Алматы, Изд-во: «Каганат», 2007. – 516 с.

Казахстанское Агентство Прикладной Экологии (КАПЭ)

ӨЗЕН ЭКОСИСТЕМАЫНДАҒЫ ТАБИҒИ ПРОЦЕСТЕРДІ БОЛЖАУДАҒЫ СТАТИСТИКАЛЫҚ ӘДІСТЕМЕЛЕР ТУРАЛЫ

Техн. ғылымд. докторы М.Ж. Бурлибаев

Өзен экосистемасының қазіргі таңда деградацияға ұшырауына байланысты өзен кешенінің табиғи суағындарының гидрологиялық тәртібін бақылаушы дәрәжесіне қалпына келтіру маңызды болып отыр. Осы мақсатта өзен экосистемасындағы антропогенттік процестердің әсерінен болатын келешектегі табиғи өзгерістерді тура болжау қажет. Сондықтан осы мақала өзен экосистемасындағы табиғи процестерді болжаудың статистикалық әдістемелеріне арналады.