Всероссийская конференция
Обработка пространственных данных в задачах
мониторинга природных и антропогенных процессов
22-25 августа 2023, Бердск, Россия

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ БОЛЬШИХ ОБЪЕМОВ ДАННЫХ ПРИ ПОСТРОЕНИИ ПРОГНОЗНЫХ ГИДРОЛОГИЧЕСКИХ МОДЕЛЕЙ

Ю.Б. Кирста, И.А. Трошкова

Институт водных и экологических проблем СО РАН

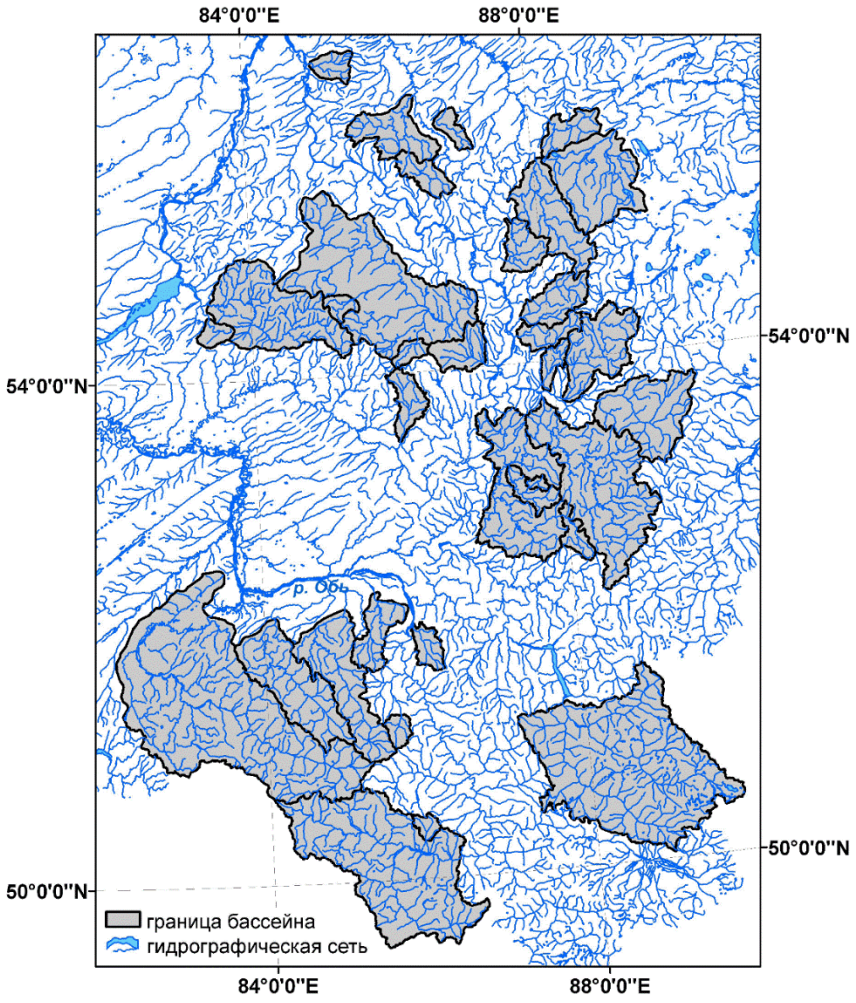
ИВЭП

Введение

- Интеллектуальный анализ данных (ИА), известный как Data Mining и Knowledge Discovery, подразумевает выделение статистически неявной зашумленной, но потенциально важной информации, содержащейся в больших выборках экспериментальных данных.
- Сейчас разработаны различные математические методы ИА, совмещающие классическую статистику, специальные технологии обработки баз данных, алгоритмы машинного обучения, искусственные нейронные сети и искусственный интеллект.
- ИА все активнее применяется в бизнесе, менеджменте, науке, здравоохранении, прогнозировании катастрофических процессов.
- Нами предложена стандартная методология ИА, обеспечивающая адекватное математическое описание сложных слабоизученных природных процессов путем извлечения скрытой информации из простых стандартных наблюдений за их динамикой.
- ИА успешно применен при моделировании агроэкосистем России и США, процессов влагообмена в почвах, прогнозе изменений климата, динамики гидрологических и гидрохимических стоков рек.

Введение

- С помощью ИА нами построена прогнозная модель пиков весенних половодий с ледоходом на 34 реках Алтае-Саянской горной страны.



- Страну отличает большое разнообразие ландшафтно-климатических зон: ледников, гольцов, горных тундр и альпийских лугов на высокогорьях, хвойных лесов, степей и полупустынь на склонах и межгорных котловинах.
- Питание многочисленных рек страны является смешанным снегово-дождевым и ледниковым. Весенняя волна половодья формируется в апреле при вскрытии рек от льда и таянии снежного покрова.

Рис. Карта-схема 34 речных бассейнов Алтае-Саянской горной страны с площадью 177–21000 км^2.

Методология ИА

- Рассмотрим методологию ИА на примере построения прогнозной модели пиков весенних половодий с ледоходом на 34 реках Алтае-Саянской горной страны.

- Методология имеет также название системно-аналитического моделирования и включает пять последовательных этапов.

- На **первом этапе** ИА определяются три иерархических уровня организации гидрологических процессов горных водосборов.

  ● *Первый наиболее высокий уровень* отвечает охватываемым территории и периоду времени. Это все 34 речных бассейна страны с ее едиными мезомасштабными метеорологическими процессами.

  ● *Второй уровень* характеризует отдельные гидрологические системы, слагающие первый уровень. Это водосборные бассейны отдельных рек.

  ● *Третий уровень* отвечает меньшим гидрологическим системам, входящим в системы второго уровня. Это отдельные ландшафты.

Методология ИА

- На **втором этапе** ИА идет подготовка данных за 1951–2020 гг. для моделирования с созданием их однородных выборок, в том числе:
  - *выбор шагов* временной и пространственной шкал описания процессов и изменения факторов среды,
  - *нормировку метеорологических факторов*, при котором наблюдаемые на 11 реперных метеостанциях среднемесячные температуры воздуха и месячные осадки нормируются на свои абсолютные среднемноголетние значения и затем усредняются по всей территории – Алтае-Саянской горной стране,
  - *нормировку пиков половодий* ($м^3/с$) на их среднемноголетние значения по каждому речному бассейну,
  - *определение границ ландшафтов* (выделено 13 типов), их площадей и высот средствами ArcGIS 10.2,
  - *нормировку* их площадей на площадь соответствующего бассейна.
- В итоговую базу данных вошли:
  - гидрологическая выборка (~300 нормированных пиков половодий,
  - две метеорологические (840 нормированных среднемесячных температур воздуха и 840 нормированных месячных осадков),
  - две ландшафтные (160 нормированных значений площадей и 160 ненормированных высот ландшафтов).

Методология ИА

- На **третьем этапе** ИА выбирается тип модели (интегро-дифференциальный, алгебраический, имитационный) и составляются различные варианты ее уравнений с соблюдением всех физических, гидрологических и других законов.
- Обязательно соотношение более 5:1 между количеством наблюдений за выходной переменной и числом параметров модели.
- Из уравнений проверяемого варианта модели создается система большой размерности, где каждое уравнение рассчитывает конкретное значение выходной переменной (пика половодий).
- Для этой системы решается обратная математическая задача с определением значений всех коэффициентов модели и ее невязки (расхождения между рассчитанными и наблюдаемыми значениями выходной переменной).
- После проверки различных вариантов описания анализируемых процессов определяется вариант с наименьшей квадратичной невязкой, который принимается за готовую математическую модель.
- Считается, что наименьшую невязку модели обеспечивает только адекватное описание всех моделируемых процессов и влияния на них факторов среды.

Методология ИА

- На **четвертом этапе** ИА оцениваются адекватность (точность) модели и ее чувствительность к вариациям факторов среды:

$$A = S_{\text{dif}}/\sqrt{2}S_{\text{obs}}$$

- где A – критерий адекватности модели, S_{dif} – стандартное (среднеквадратическое) отклонение у невязки модели, S_{obs} – стандартное отклонение у наблюдаемых данных.
- Интервал A=0–0.71 отвечает степени совпадения расчетных и наблюдаемых значений выходной переменной и их совпадению при A~0. A подобен критерию Нэша-Сатклиффа $\text{NSE} = 1–2A^2$.

$$FS = (A')^2 - (A)^2 = \frac{(S'_{\text{dif}})^2 - (S_{\text{dif}})^2}{2(S_{\text{obs}})^2} = \frac{2(S_{\text{fac}})^2}{2(S_{\text{obs}})^2} = \frac{(S_{\text{fac}})^2}{(S_{\text{obs}})^2}$$

- где FS – чувствительность к входному фактору, A' – величина A, полученная после случайного перепутывания значений фактора, $(S_{\text{dif}})^2$ – дисперсия невязки модели, $(S'_{\text{dif}})^2$ – эта же дисперсия после перепутывания фактора, $(S_{\text{fac}})^2$ – вклад природных вариаций фактора в дисперсию выходной переменной $(S_{\text{obs}})^2$.

Методология ИА

- **На пятом этапе** ИА оценивается качество (A, NSE) упрощенного *прикладного варианта* модели, описывающего конкретный объект.
- Значения наиболее важных параметров варианта обновляются по данным об этом объекте через новое решение обратной задачи.
- У аналитических моделей сложноорганизованных природных систем дисперсия невязки выходной переменной слагается из компонентов, обусловленных погрешностями наблюдений за факторами среды и погрешностью самих уравнений модели:

$$2A^2 = (S_{\text{dif}})^2/(S_{\text{obs}})^2 \approx \sum_i FS_i + \sum_j FS_j \times 2A_j^2 + 2A_0^2$$

- где $\sum_i$, $\sum_j$ – суммирование по факторам i и j, $i \neq j$; FS_i – вклад в $(S_{\text{dif}})^2$ от вариаций входного фактора i (равен чувствительности к фактору i, отсутствующего в прикладной модели; FS_j – вклад в $(S_{\text{dif}})^2$ от вариаций входного фактора j (равен чувствительности модели к фактору j), который по-прежнему учитывается в прикладной модели; A_j^2 – доля в дисперсии наблюдаемых значений входного фактора j, формируемая ошибками его наблюдений; A_0^2 – компонент дисперсии $(S_{\text{dif}})^2$ с искомой адекватностью A_0 прикладного варианта модели.

Прогнозная модель пиков весенних половодий

- В рамках ИА разработана модель для среднесрочного прогноза пиков весенних половодий с минимальной невязкой прогнозов:

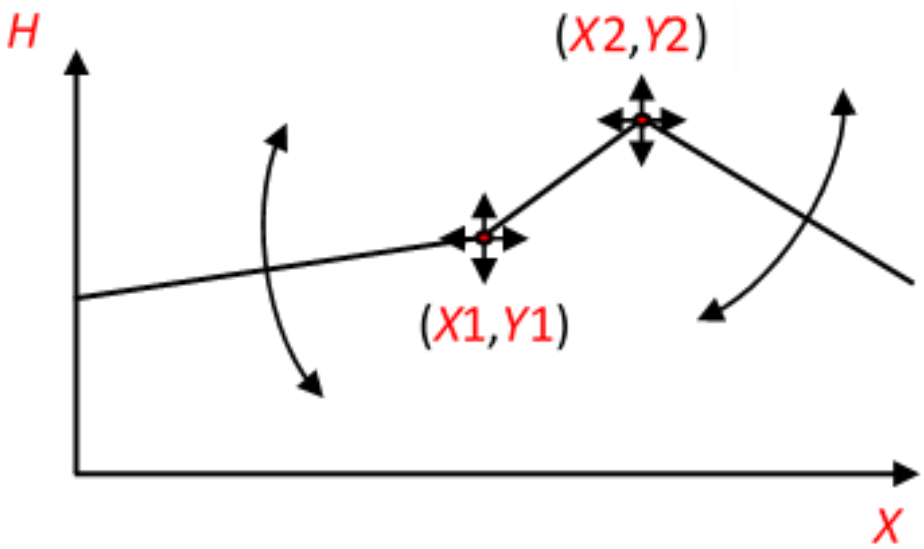
$$Q^i = H(c_1, c_2, 1{,}1, c_3, c_4, P_1)\left\{\sum_k a_k S_k^i P_1 H\left(c_9, c_{10}, 1{,}1, c_{11}, c_{12}, h_k^i\right)\right.$$
$$\left. + \sum_k b_k S_k^i P_2 H(c_5, c_6, 1{,}1, c_7, c_8, T_2) H\left(c_9, c_{10}, 1{,}1, c_{11}, c_{12}, h_k^i\right)\right\} + d$$

- где Q^i – ежегодно прогнозируемое значение пика половодья для выходного створа речного бассейна *i*, *i*=1-34; первое и второе слагаемые в правой части уравнения соответствуют вкладам предшествующих осенних (IX–XI) и зимних (XII–III) месяцев; a_k, b_k – параметры, характеризующие вклад *k*-го ландшафта за осенние и зимние месяцы, *k*=1-13; S_k^i – нормированная площадь *k*-го ландшафта бассейна *i*; h_k^i – высота ландшафта, метры н.у.м.; P_1, P_2 – отклонения нормированных месячных осадков от их среднемноголетних значений в среднем за осенние и зимние месяцы соответственно; T_2 – отклонение нормированных среднемесячных температур воздуха от среднемноголетнего значения за зимние месяцы, $c_{1\text{-}4}$, $c_{5\text{-}8}$, $c_{9\text{-}12}$ – параметры, d – потери талых вод на промачивание почв и просачивание в трещиноватые горные породы.

Прогнозная модель пиков весенних половодий

- кусочно-линейная функция H с изменяемыми значениями параметров позволяет аппроксимировать широкий спектр зависимостей между процессами и факторами среды:

$$H(X1,X2,Y1,Y2,Z1,Z2,X)=\begin{cases} Y1+Z1(X-X1), & \text{если } X<X1 \\ \dfrac{Y2-Y1}{X2-X1}(X-X1)+Y1, & \text{если } \begin{cases} X1\le X<X2 \\ X1\neq X2 \end{cases} \\ Y2+Z2(X-X2), & \text{если } X\ge X2 \end{cases}$$



$X1, X2, Y1, Y2, Z1, Z2$ – параметры, определяемые в ходе решения обратной математической задачи, X – меняющийся входной фактор или переменная модели.

Прогнозная модель пиков весенних половодий

- В правой части уравнения модели суммируются 13+13=26 различных гидрологических процессов, которые формируют пики весеннего половодья в каждом из 34 речных бассейнов. Одновременный анализ этих процессов при решении обратной математической задачи представляет собой собственно **ИА** гидрологических данных.
- Значения пиков половодий (~300) подставлены в систему уравнений вместо Q^i для конкретных бассейнов и лет за период 1951–2020 гг.
- Система решена оптимизационными средствами MATLAB, в результате чего определены 39 параметров модели и ее невязка.
- Верификация модели на произвольно исключаемых бассейнах показала, что их невязки для среднем не отличались от получаемой для остальных 33 бассейнов, используемых при идентификации, и формально подтвердила универсальность разработанной модели.
- Чувствительность прогнозной модели пиков половодий к вариациям осенне-зимних осадков составила *FS* = 29% от дисперсии $(S_{\text{obs}})^2$, ландшафтной структуры – 14%, зимних температур воздуха – 0.8% и высоты ландшафтов – 0.1%. Дополнительная независимая оценка чувствительности пиков половодий к весенним (апрельским) осадкам и температурам (из прогноза исключены) дала для них 22% и 6%.

Прогнозная модель пиков весенних половодий

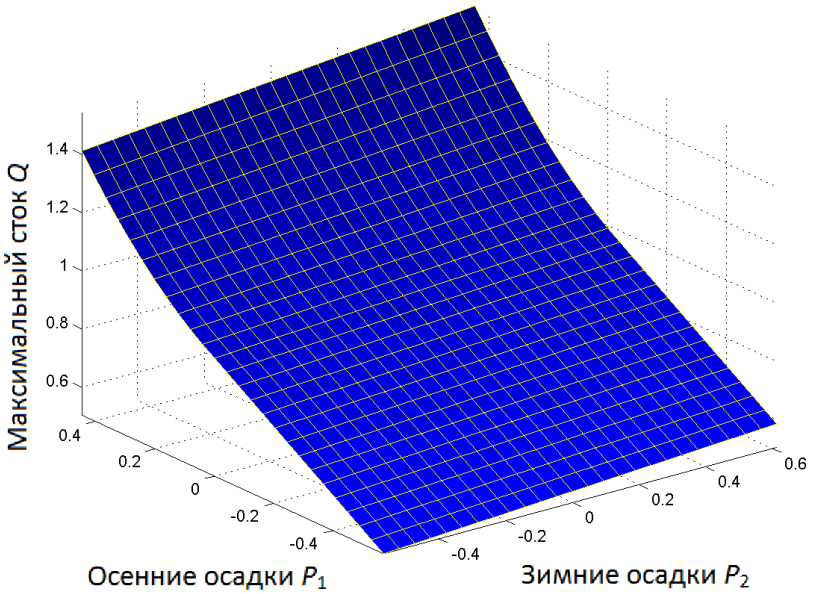
- На последнем пятом этапе ИА оценивается адекватность (точность) прикладного варианта модели при прогнозе пиков половодий в любом конкретном речном бассейне.
- У прикладной модели отсутствуют изменения площадей, высот и гидрологических характеристик ландшафтов, то есть эти факторы могут быть из нее убраны. Тогда модель принимает вид:

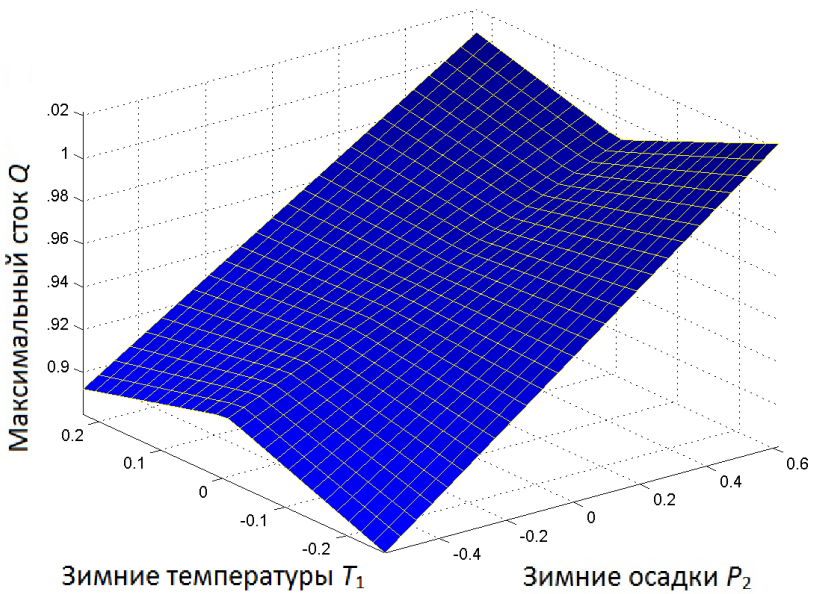
$$Q = H(c_1, c_2, 1{,}1, c_3, c_4, P_1)\{aP_1 + bP_2 H(c_5, c_6, 1{,}1, c_7, c_8, T_2)\} + d$$

- где $c_{1\text{-}4}$, $c_{5\text{-}8}$ остаются прежними, а значения трех параметров a, b, d обновляются по данным выбранного бассейна с помощью решения обратной задачи или обычного регрессионного анализа.
- Используя ранее полученные значения адекватности A для исходной модели, нормированных осенне-зимних осадков, зимних температур воздуха, а также найденные значения чувствительности к факторам, легко рассчитываем адекватность прикладной модели A_0=0.33.
- Полученные значения A_0=0.33 и NSE_0=1–2$(A_0)^2$ =0.78, характеризующие точность среднесрочных прогнозов пиков весенних половодий, отвечают высокому качеству гидрологических моделей (0.75<NSE≤1.0), что особенно значимо для половодий горных рек.

Прогнозная модель пиков весенних половодий

- Практическое отсутствие чувствительности пиков половодий (0.1%) к высоте h_k^i подтверждает адекватность описания динамики метеорологических факторов через их нормированные месячные значения, не зависящие от координат и высоты местности.
- На рисунке приведены зависимости прогнозируемых пиков от метеоусловий зимы и предшествующей осени, осадки которой в разной степени увлажняют уходящие в зиму почвы и опосредованно влияют на интенсивность весеннего половодья.





ИВЭП


Спасибо за внимание !
Финансовая поддержка РНФ – грант 22-27-00058
Lake Shavlinskoe, r. Shavla, North Chu ridge

ИА и прогнозная модель пиков весенних половодий

- Y.B. Kirsta, I.A. Troshkova, 2023. High-performance forecasting of spring flood in mountain river basins with complex landscape structure. *Water*, 15(6): 1080 https://doi.org/10.3390/w15061080 https://www.mdpi.com/2073-4441/15/6/1080

- Y.B. Kirsta, I.A. Troshkova, 2023. Deep process-data mining for building of analytical models: 1. Medium-term forecast of spring flood extremes for mountain rivers. *Eurasian Journal of Mathematical and Computer Applications*, 11(3). Принята в печать.

- Y.B. Kirsta, I.A. Troshkova, 2023. Deep process-data mining for building of analytical models: 2. Influence of winter-spring temperatures and precipitation on spring flood extremes for mountain rivers. *Eurasian Journal of Mathematical and Computer Applications*, 11(4). Принята в печать.