УДК 004.89

© 2014 г. **И.Л. Артемьева**, д-р техн. наук
(Дальневосточный федеральный университет,
Институт прикладной математики ДВО РАН, Владивосток),
Б.И. Гарцман, д-р геогр. наук
(Тихоокеанский институт географии ДВО РАН, Владивосток),
А.А. Ходеев,
А.А. Коваленко,
(Дальневосточный федеральный университет, Владивосток)

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ДЛЯ ЗАДАЧИ РАЗДЕЛЕНИЯ ГИДРОГРАФА РЕЧНОГО СТОКА ПО ИСТОЧНИКАМ ПИТАНИЯ

Рассматривается задача разработки автоматизированной системы разделения гидрографа стока по генетическим составляющим. Описывается последовательность этапов для нахождения потенциальных источников. Построена и описана математическая модель для этапа грубого отсева.
Ключевые слова: модель онтологии, основанная на онтологии программная система, нахождение потенциальных источников формирования речного стока.

Введение

Одной из наименее изученных прикладных задач, с которыми имеет дело гидролог, является задача нахождения потенциальных источников, из которых формируется речной сток. Существует методика, основанная на использовании природных химических трассеров в рамках модели смешения (модель EMMA [1]), позволяющая разделить гидрограф стока по источникам питания. Данная методика предполагает анализ большого количества проб для получения достоверных результатов. Нахождение потенциальных источников выполняется в несколько этапов, среди которых: этап грубого отсева, этап стандартизации данных, этап нахождения главных компонент [5], этап разделения гидрографа стока и другие. На текущий момент существуют программные системы, с помощью которых специалисты предметной области решают поставленную задачу.

CUASHI содержит в себе большинство методов, необходимых для решения задачи разделения гидрографа стока по источникам питания. Однако в силу своей многофункциональности данная система крайне сложна для использования. Кроме того, адаптация и сопровождение системы требуют дополнительных затрат на обучение пользователей для работы с системой [6].

Пакет прикладных программ «Гидролог» представляет собой достаточно удобный механизм решения задач, так как имеет привычный табличный интерфейс, но, к сожалению, содержит в себе методы, не позволяющие в полной мере решить задачу разделения гидрографа стока по источникам питания [8].

SYSTAT – это набор типовых методов статистического анализа: описательная статистика, дисперсионный, корреляционный и спектральный анализы, сглаживание, прогнозирование, пошаговая и нелинейная регрессия, кластерный и факторный анализ и т. д. С помощью данного пакета можно повысить качество работ и унифицировать представление результатов. Но, к сожалению, задачу определения генетической структуры речного стока в SYSTAT решить нельзя, так как в данной программе не реализован самый главный метод EMMA-анализа – метод главных компонентов [9].

Программа STATGRAPHICS – инструмент статистического анализа. STATGRAPHICS имеет дружественный интерфейс и тщательно подготовленную документацию, которые способствуют быстрому освоению пакета как специалистами в области математической статистики, так и представителями других сфер деятельности [7]. С помощью STATGRAPHICS можно решить задачу определения генетических источников, но только с помощью специалиста: программа не может отобрать сама рабочие варианты, не может сформировать наборы данных, все это приходится делать специалисту вручную.

Каждая из вышеперечисленных программ не позволяет полностью решать задачу разделения гидрографа стока по источникам питания: одни системы обладают недостаточным для решения задачи функционалом, другие имеют высокую степень сложности в изучении и сопровождении. В связи с этим возникает необходимость в автоматизированной интеллектуальной программной системе для анализа и статистической обработки данных, которая позволяла бы специалисту предметной области получать результаты обработки в упрощенном виде (в виде графиков), накапливать знания, результаты анализа имеющихся данных, а также сохранять результаты выполнения методики.

Цель данной статьи – формальное описание методики, используемой специалистами предметной области гидрологии для решения задачи разделения гидрографа стока по генетическим составляющим.

Описание программной системы

Система рассчитана на работу с двумя типами пользователей: инженер-специалист, он же специалист предметной области, который непосредственно работает с данными (добавляет в систему пробы, редактирует имеющийся набор проб), а также решает задачи обработки данных, и эксперт, который имеет возможность работать со знаниями, заложенными в программную систему (рис. 1).

Разрабатываемая программная система в итоге должна решать следующие задачи. Первая – задача автоматизации научных исследований и накопление полученных знаний. Так как в России исследование с применением модели EMMA проводится впервые [1], необходимы адаптация и отработка методики. Для решения этой задачи необходим оператор системы, в качестве которого выступает

специалист предметной области. Использование системы специалистом намного ускорит ход исследования.

Следующая задача, которую можно решить с использованием интеллектуальной программной системы, – задача мониторинга. В перспективе предполагается, что система сама должна решать повседневную задачу мониторинга, без участия оператора. Ниже опишем основные компоненты программной системы.

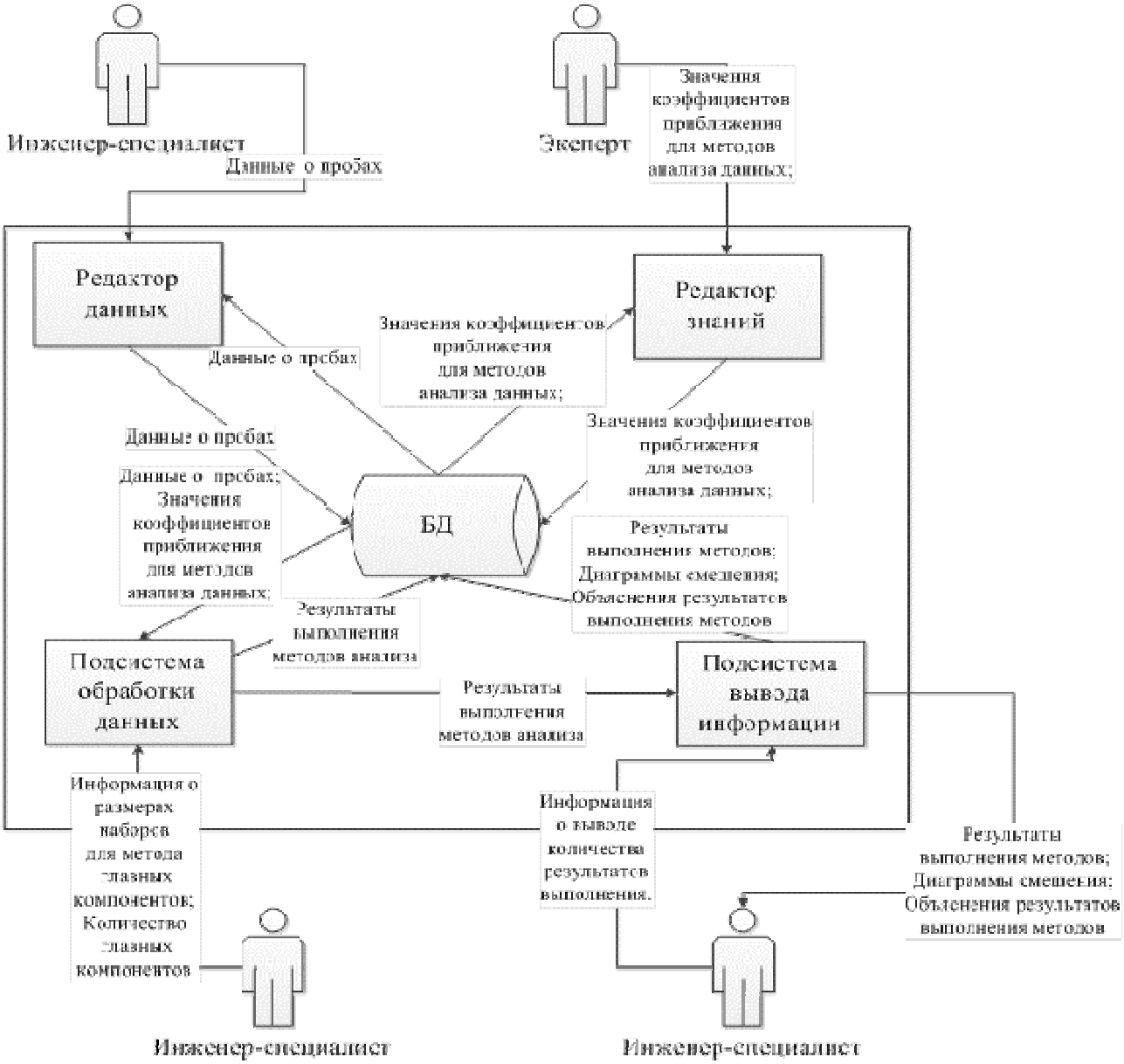

Рис.1. АКД программной системы.

Редактор данных

Так как методика с использованием модели ЕММА предполагает работу с большим объемом данных (пробами воды), в программной системе должна быть реализована подсистема, которая позволила бы заносить новые массивы данных в систему и обрабатывать уже существующие (редактор данных).

Инженер-специалист может внести данные о пробах в подсистему двумя способами: вручную или же импортировать существующие данные из excel-файла (рис. 2).

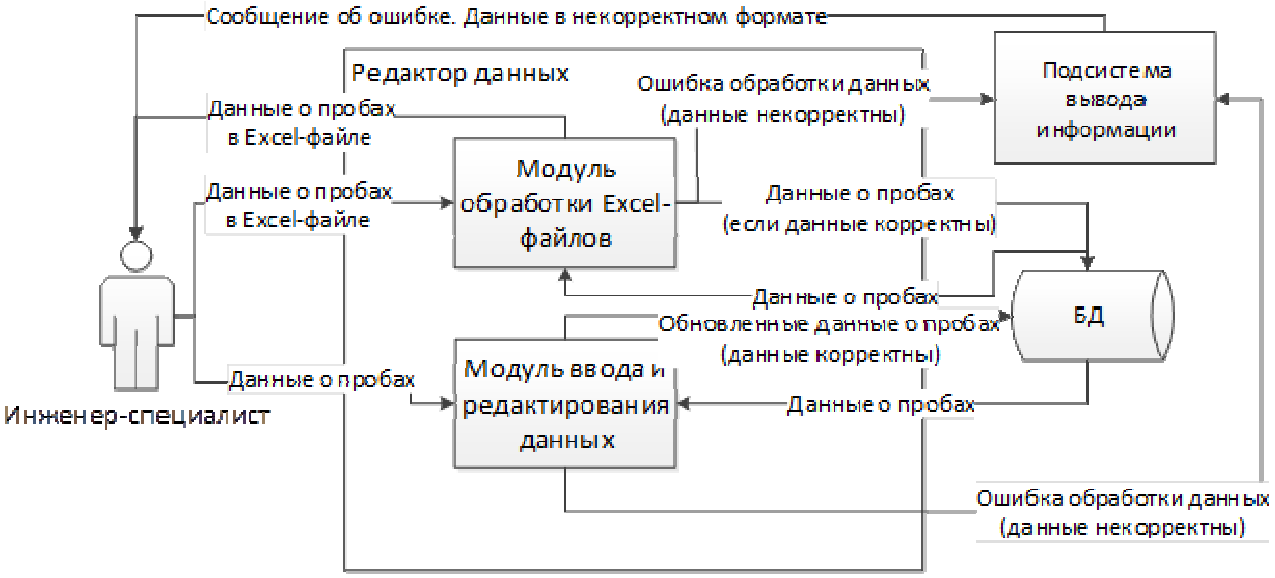

Рис. 2. Структура подсистемы редактирования данных.

В первом случае будет задействован модуль ввода и редактирования данных, который проверит предоставленные на вход данные на соответствие правилам, заложенным в систему. Правила задаются моделью онтологии [4], фрагмент которой приведен ниже.

1. Сорт Идентификаторы проб: {}N \ ∅

В предметной области существует непустое множество проб. Каждая проба имеет свой уникальный идентификатор.

2. Сорт Дата отбора пробы: Идентификаторы проб → Дата
3. Сорт Место отбора пробы: Идентификаторы проб → Места отбора
4. Сорт Элементы пробы: Идентификаторы проб →{}Компоненты

У каждой пробы имеется некоторый набор элементов.

Ограничения

5. (i: Компоненты, s1: Идентификаторы проб, s2: Идентификаторы проб) (s1 ≠ s2) & (i ∈ Элементы пробы(s1)) => i ∈ Элементы пробы(s2)

Для всех проб набор компонентов одинаковый. Если компонент i присутствует в пробе s1 и отсутствует в пробе s2, то концентрация компонента в пробе s2 равна нулю.

6. (v1: Идентификаторы проб, v2: Идентификаторы проб) Дата отбора пробы(v1) ≠ Дата отбора пробы(v2) => v1 ≠ v2

Для каждой пробы определена своя дата взятия пробы.

7. (v1: Идентификаторы проб, v2: Идентификаторы проб) Место отбора пробы(v1) ≠ Место отбора пробы(v2) => v1 ≠ v2

Для каждой пробы определено свое место взятия.

8. Сорт Определяемые компоненты пробы: Идентификаторы проб →{} Компоненты
9. (i: Идентификаторы пробы) Определяемые компоненты пробы(i) ⊆ Элементы пробы(i)
10. Сорт Концентрация: (i: Идентификаторы проб, e: Определяемые компоненты пробы(i)) → R[0; ∞)

Концентрация – это «вес» элемента в конкретной пробе. Если концентрация элемента в пробе равна нулю, то считается. Что этот элемент в данной пробе отсутствует

11. Количество элементов пробы ≡ (λ(i: Идентификаторы пробы) μ(Элементы пробы(i))

Количество элементов пробы – это число наименований всех элементов, обнаруженных в пробе.

12. Количество проб в системе ≡μ(Идентификаторы проб)

В системе конечное количество проб

13. Сорт Тип пробы: (s: Идентификаторы проб) → Возможные типы проб

Тип пробы – это категория конкретной пробы: категорией может быть осадки, грунтовые, русловые и т.д.

Данная подсистема также позволяет осуществить импорт данных из excel-файлов (с расширениями .xls, .xlsx) в базу данных системы. Если в качестве входных данных редактором данных был получен excel-файл, он будет направлен на вход модулю обработки excel-файлов. Данный модуль производит синтаксический анализ данных и проверку полученной структуры на соответствие правилам системы, которая, в свою очередь, также задается моделью онтологии. Если данные корректны, подсистема сохранит новую информацию в базе, иначе сообщит системе вывода информации о некорректности введенных данных, а та, в свою очередь, выдаст соответствующее сообщение специалисту.

Необходимость добавления в систему модуля обработки excel-файлов заключается в следующем: специалисты предметной области предпочитают хранить всю информацию о пробах в файлах Microsoft Excel, переносить все данные в систему вручную крайне затруднительно. Следовательно, возникает потребность создания некоторого модуля работы с excel-файлами, который позволил бы специалисту добавлять, редактировать имеющиеся данные, а также по желанию экспортировать хранящуюся в базе данных информацию обратно в excel-файл.

Редактор знаний

Каждая проба обладает своим набором компонентов, своим местом взятия и принадлежит к определенному типу водной массы. Однако допустимое множество компонентов, множество мест взятия и типы водных масс для всех проб одинаковы. Указанные сведения являются знаниями предметной области гидрологии суши.

Правила задаются моделью онтологии, фрагмент которой приведен ниже (модель онтологии представлена с использованием языка, описанного в [4]).

1. Сорт Компоненты: {}N\ ⌀

В предметной области существует непустое множество компонентов; каждый из компонентов имеет название.

2. Сорт Тип компонента: Компоненты → {макрокомпоненты, микрокомпоненты, физические характеристики}
3. Сорт Места отбора: {}N \ ∅
4. Дата ≡ (x День, Месяц, Год, Час, Минута)
5. День ≡ I[1; 31]
6. Месяц ≡ {Январь, Февраль, Март, Апрель, Май, Июнь, Июль, Август, Сентябрь, Октябрь, Ноябрь, Декабрь}
7. Год ≡ I[начальный год; конечный год]
8. Час ≡ I[0; 23]
9. Минута ≡ I[0; 59]
10. Сорт начальный год: I[1900; ∞)
11. Сорт конечный год: I(начальный год; ∞)
12. Сорт Приоритет: Компоненты →I[0;∞)

Приоритет – это число в диапазоне [0, 5], выставляемое экспертом каждому элементу пробы, обозначающее важность элемента. Для всех проб приоритет для некоторого элемента один и тот же.

13. Сорт Ошибка измерения: Компоненты →R[0, +∞]
14. Сорт Комментарий: Идентификаторы пробы →N

К каждой пробе может быть комментарий – краткая заметка об особенностях условий, в которых была взята данная проба. Например: «Шел дождь».

15. Сорт Широта: Места отбора → R[0,90]
16. Сорт Долгота: Места отбора → R[0,180]
17. Сорт Высота взятия: Места отбора →R
18. Сорт Минимально допустимая концентрация: R[0;1]
19. Сорт Минимально допустимый процент для элемента при недостаточности данных: R[0; 100]

Элементы после недостаточности данных – это множество компонентов, которые содержатся в достаточном для дальнейшего анализа количестве проб.

20. Сорт Возможные типы проб: {}N \ ∅

Возможные типы проб содержат в себе категории проб. Возможными типами проб могут являться: осадки, грунтовые воды.

21. Сорт Анализируемый тип проб: Возможные типы проб

Анализируемый тип пробы – это единственная категория, для которой будет проводиться метод главных компонент, а также строиться модель смешения. В частности, это русловые воды.

Данная подсистема позволяет вносить новые и редактировать существующие знания предметной области (рис. 3).

Рис. 3. Структура подсистемы редактирования знаний.

Подсистема обработки данных

Как было указано выше, базовой задачей является определение источников поступления воды в речных бассейнах. Подсистема обработки данных (рис. 4) решает поставленную задачу с помощью следующих содержащихся в ней модулей: модуля грубого отсева, цель которого – устранение «ненужных» данных, модуля стандартизации данных, в результате которого происходит приведение зна-

чений концентраций компонентов проб к единому формату и размерности и модуля главных компонент.

Рис. 4. Структура подсистемы обработки данных.

Модуль грубого отсева

Перед выполнением анализа эксперту в "редакторе знаний" следует выделить из имеющегося набора типов водных масс проб анализируемый тип проб. Отметим, что каждая проба воды имеет характеристику «Тип водных масс» – это отметка пробы, которая указывает, из какого типа природных вод отобрана проба (например: атмосферные воды, подземные воды). Только те пробы, тип которых совпадает с анализируемым типом, будут поданы на вход модулю грубого отсева.

Данный модуль содержит в себе три подмодуля: модуль проверки на пропуски данных, проверки данных в соответствии с приборной ошибкой измерения и проверки данных на корреляцию. В результате выполнения модуля грубого отсева некоторые компоненты будут исключены из дальнейшего анализа.

При грубом отсеве вначале выполняется проверка на пропуски данных. Этот этап позволяет «отсеять» из набора проб те компоненты, которые отсутствуют в большинстве проб. Правило, по которому некоторый компонент отбрасывается из дальнейшей обработки: компонент отсутствует более чем в N% проб.

Компонент считается отсутствующим в пробе, если его концентрация в данной пробе меньше минимально допустимой. Минимально допустимую концентрацию, а также величину N определяет эксперт предметной области.

1. Пробы с нулевой концентрацией ≡ (λ (x: Компоненты) {i ∈ Идентификаторы проб | Концентрация(i, x) < Минимально допустимая концентрация})
2. (x: Компоненты) Процент для отсеивания при недостаточности(x) ≡ μ(Пробы с нулевой

концентрацией(x)) / Количество проб в системе

3. Элементы после недостаточности данных ≡ {(x ∈ компоненты) | (Процент для отсеивания при недостаточности(x) > Минимально допустимый процент для элемента при недостаточности данных)}

Вслед за этапом проверки на пропуски данных следует проверка данных в соответствии с приборной ошибкой измерения. Этот этап позволяет отобрать из набора проб недостоверные компоненты.

1. Элементы после проверки на ошибку $\equiv U_{i \in \text{Идентификаторы проб}}$ {

(x: Элементы после недостаточности данных) |

Если x ∈ Элементы пробы(i), то Концентрация(i,x)>2*Ошибка измерения(x), иначе Концентрация(i, x)= 0}

Элементы после проверки – это множество компонентов, которые удовлетворяют следующему условию: если компонент присутствует в пробе, то концентрация элемента в пробе превышает в 2 раза приборную ошибку измерения для данного элемент, если компонент не присутствует в пробе, то его концентрация равна 0.

Концентрация компонента в пробе является достоверной, если она превышает приборную ошибку измерения данного компонента больше чем в два раза. Правило, по которому некоторый компонент отбрасывается из дальнейшей обработки: процент недостоверных концентраций данного компонента в наборе проб меньше минимально допустимого процента M. Величину M определяет эксперт предметной области.

И наконец, проверка данных на корреляцию [2]. На данном этапе происходит проверка на корреляцию компонентов друг с другом.

1. Учитываемые элементы ≡ $U_{i \in \text{Элементы после проверки на ошибку}}$ {(x∈ Элементы после проверки на ошибку) | i ≠ x & |коэффициент корреляции(x, i)| * 100 < Максимально допустимый процент коэффициента корреляции для элементов}

Учитываемые элементы – это множество компонентов, прошедших этап грубого отсева.

Модуль стандартизации данных

Пробы, прошедшие этап грубого отсева, поступают на вход модулю стандартизации в виде некой «очищенной матрицы» концентраций компонентов проб. Очищенная матрица – матрица, строками которой являются пробы; столбцы – компоненты, содержащиеся в пробах; значения – концентрации компонентов в пробе. Процесс стандартизации происходит следующим образом: для каждого компонента определяется его среднее значение и среднеквадратичное отклонение на основании данных из очищенной матрицы, после чего пересчитываются концентрации компонента в каждой пробе по следующей формуле (1), где x_I – текущая концентрация компонента в *i*-й пробе; X – среднее значение для компонента; σ – среднеквадратичное отклонение для компонента; z_i – стандартизированная концентрация компонента в *i*-й пробе. Стандартизация проводится для всех типов проб в совокупности.

1. Сорт Пробы для элементов: Компоненты → {}Идентификаторы проб

Ограничения

2. (v1: компоненты) (v2: пробы для элементов (v1)) v1 → учитываемые элементы

Какой бы ни был элемент, проба входит в множество проб для этого элемента, если этот элемент является учитываемым элементом.

3. (v1: компоненты) Число проб для компонента (v1) ≡ μ (Пробы для элементов (v1))
4. Сорт Дисперсия компонента: компоненты →R(-∞; +∞)
5. (*v2* учитываемые элементы)

Дисперсия компонента(*v2*) = (1 / число проб для компонента(*v2*)) * ∑(i ∈ [1; число проб для компонента(*v2*)])(концентрация(i, *v2*)-среднее значение для компонента(*v2*))^2

Дисперсия учитываемого компонента вычисляется по следующей формуле: $\delta_x = 1 / n * \sum^n (x_i - x_{cp})^2$*, где n – число проб для компонента,* x_i *– концентрация в i-ой пробе,* x_{cp} *– среднее значение для компонента.*

6. Сорт стандартизированная концентрация: (п→ пробы, э→ учитываемые элементы(п)) →R[0; ∞)

Стандартизированная концентрация – это новое значение элемента в конкретной пробе, полученное в результате проведения этапа стандартизации данных.

7. *(v1:* пробы) (*v2:* учитываемые элементы) Стандартизированная концентрация(*v1, v2*) = (1 / Дисперсия компонента(*v2*))* (Концентрация(*v1, v2*) – Среднее значение компонента (*v2*))

Подсистема вывода информации

Помимо представленных выше подсистем, в системе присутствует подсистема (рис. 5), которая генерирует объяснения и обоснование полученных результатов анализа и обработки проб (модуль объяснения).

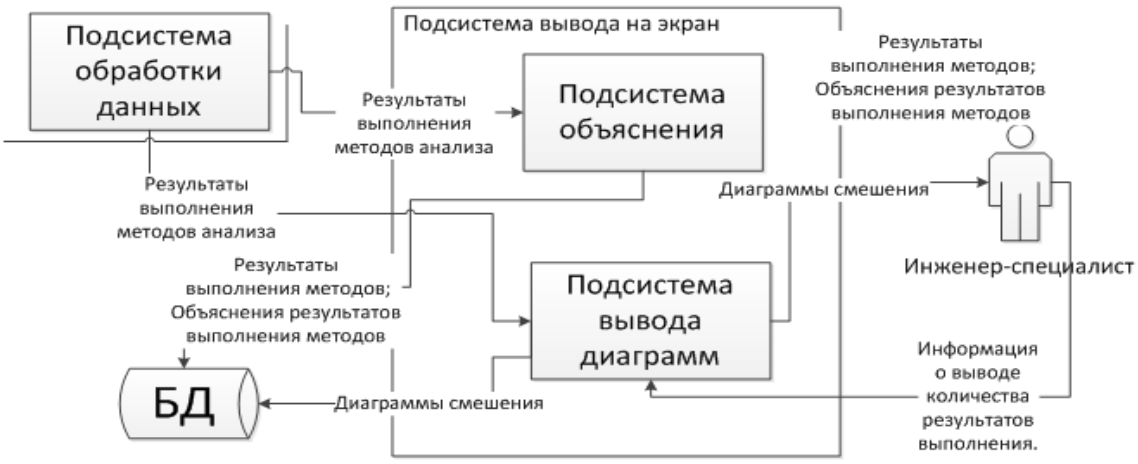

Рис. 5. Структура подсистемы вывода информации.

Помимо модуля формирования пояснений, в подсистеме присутствует модуль вывода диаграмм. Модуль вывода диаграмм создает графическое представление результатов этой обработки. Оба модуля получают на вход результаты выполнения методов анализа из подсистемы обработки данных.

1. Пробы для источника ≡ (λ (v: Возможные типы проб) :{x ∈ Идентификаторы проб | Тип пробы (x) = v})

Пробы для источника – это такие пробы, которые принадлежат указанному источнику.

2. Сорт Тип источника: Возможные типы проб \ {Анализируемый тип проб}

Типом источника является множество типов проб, за исключением анализируемого типа проб.

3. Среднее для компонента типа источника ≡ (λ (v: Тип источника, п: пробы для источника(v), q: учитываемые элементы)) ($\sum_{(i \in \text{Пробы для источника(v)})}$ стандартизированная концентрация(i, q)) / μ(Пробы для источника (v))

Среднее для компонента типа источника – это функция, которая определяет среднее значение стандартизированной концентрации компонента для пробы, принадлежащей указанному типу источника.

4. Сорт Все вершины: {}N

Все вершины – это всевозможные отметки на графике.

5. Сорт Название осиX: (i: номер набора) π(1, Подходящие наборы(i))
6. Сорт Название осиY: (i: номер набора) π(2, Подходящие наборы(i))
7. Сорт Отметка пробы: (v → Пробы для анализа, i → Номер набора) → <стандартизированная концентрация(v, Название оси X(i)), стандартизированная концентрация(v, Название оси Y(i)))>

Отметка пробы – отображение пробы на графике. Отметка пробы сопоставляет точку на графике значениям стандартизированных концентраций двух компонентов, которые являются осями координат для указанного номера набора.

8. Сорт Вершина треугольника: (t → Тип источника, s → Пробы для источника(t), i → Номер набора) → <Среднее для компонента типа источника(t, s, Название оси X(i)), Среднее для компонента типа источника(t, s, Название оси Y(i))>

Вершины треугольника – это всевозможные вершины треугольника на графике. Каждая вершина треугольника – это вершина, координатами которой являются средние значения для компонентов.

Заключение

В работе представлено описание интеллектуальной программной системы, позволяющей разделить гидрограф речного стока по источникам питания, описаны этапы грубого отсева и стандартизации данных, а также построена математическая модель для этих этапов.

ЛИТЕРАТУРА

1. *Губарева Т.С., Гарцман Б.И., Шамов В.В., Болдескул А.Г., Кожевникова Н.К.* Экспериментальные исследования генетической структуры стока с помощью химических трассеров: постановка задачи // Инженерные изыскания. – 2013. – № 1. – С. 60-69.
2. *Лакин Г.Ф.* Биометрия. – М.: Высшая школа, 1990.
3. *Шитиков В.К., Розенберг Г.С., Костина Н.В.* Методы синтетического картографирования территории // Количественные методы экологии и гидробиологии. – Тольятти: СамНЦ РАН. – 2005. – С. 167-227.
4. *Клещев А.С., Артемьева И.Л.* Необогащенные системы логических соотношений // Научно-техническая информация, серия 2. – 2000. – № 7. – С. 18-28; № 8. – С. 8-18.
5. *Померанцев А.* Метод Главных Компонент (PCA) // Российское хемометрическое общество. [Электронный ресурс]. – Режим доступа: http://rcs.chemometrics.ru/Tutorials/pca.htm (дата обращения: 23.12.2013).
6. *CUAHSI* WaterOneFlow Workbook (version 1.1). A guide to using CUAHSI's WaterOneFlow web services to retrieve hydrologic time series data. 2010 г. [Электронный ресурс]. – Режим доступа: http://his.cuahsi.org/documents/HISDoc5_UseWebServices11.pdf.
7. *Дюк В.А.* Обработка данных на ПК в примерах. – СПб.: Питер, 1997.
8. *Волчек А. А., Парфомук С. И.* Пакет прикладных программ для определения расчетных характеристик речного стока // Вестник Полесского государственного университета. Серия природоведческих наук. – 2009. – № 1. – С. 22-30.
9. *Тюрин Ю.Н., Макаров Д.А.* Статистический анализ данных на компьютере / под ред.В.Э. Фигурнова. — М.: ИНФРА-М, 1998. — С. 528.

Статья представлена к публикации членом редколлегии М.А. Гузевым.

E-mail:
Артемьева И.Л. – iartemeva@mail.ru;
Гарцман Б.И. – gartsman@inbox.ru;
Ходеев А.А. – artemhodeev@bk.ru;
Коваленко А.А. – kovalenko.staiya@gmail.com.