В.А. Кузьмин, А.Г. Сурков, К.В. Шеманаев

ПРИНЦИПЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ В АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ ПРОГНОЗИРОВАНИЯ СТОКА

V.A. Kuzmin, A.G. Surkov, K.V. Shemanaev

PRINCIPLES OF AUTOMATIC DATA PROCESSING IN AUTOMATED STREAMFLOW FORECASTING SYSTEMS

Рассмотрены основные принципы полностью автоматизированной обработки данных гидрометеорологических наблюдений, используемых для прогнозирования речного стока автоматизированными системами. Методические отличия между автоматической и «ручной» обработкой данных показаны с учетом рекомендаций ВМО и международных стандартов. Показаны возможные пути развития автоматизированных систем прогнозирования различных геофизических процессов и явлений в Российской Федерации.

Ключевые слова: речной сток, автоматизированное прогнозирование, прогностическая система, автоматическая обработка данных

Basic principles of fully automated processing of hydrometeorological data used in automated streamflow forecasting systems are considered. Presented methodological differences between automatic and manual data processing are based on WMO recommendations and international standards. Possible ways of further development of automated forecasting geophysical systems in the Russian Federation are shown.

Key words: river runoff, automatic forecasting, forecasting system, automatic data processing.

Общие положения

В настоящее время Российским государственным гидрометеорологическим университетом выполняется целый ряд научно-исследовательских работ, подразумевающих разработку автоматизированных систем прогнозирования (АСП) различных гидрологических и метеорологических процессов и явлений.

Для последующего внедрения и дальнейшей коммерциализации разрабатываемых АСП необходимо обеспечить научно и технически обоснованные гарантии качества, главным образом основанные на соблюдении установленных стандартов в организации функционирования и взаимодействия всех подсистем АСП [1-3], включая подсистему контроля качества и обработки поступающих данных автоматических гидрометеорологических измерений, являющуюся важнейшим элементом любой АСП. В то же время, в связи с тем, что автоматизированные системы гидрологического прогнозирования во всем мире разрабатываются и внедряются лишь со второй половины 70-х годов прошлого века, национальными гидрометеорологическими службами до сих пор практически не разрабатывались и, следовательно, не используются единые стандарты полностью автоматизированной обработки данных.

Во многих случаях используются регламентированные действующими нормативными документами принципы и методы контроля качества и обработки данных, применяемые персоналом прогностических подразделений либо в «ручном», либо в полуавтоматическом режиме (то есть, под контролем оператора).

Полностью автоматизированное (например, фоновое) прогнозирование речного стока подразумевает появление человека лишь на заключительном этапе, когда автоматически выпущенный прогноз (в любой форме, например, в виде карты зон повышенного риска наводнения), построенный на основе автоматически собранных и обработанных данных, появляется перед лицом, ответственным за принятие решения на его основе. Многих привычных этапов процедуры сбора и обработки данных теперь не существует; исчезли бумажные журналы наблюдений; компьютерное архивирование данных автоматических измерений стало стандартной практикой в большинстве стран, поэтому обработка данных подразумевает их автоматическое конвертирование в нужный формат на раннем этапе (например, в момент, когда измерения поступает в электронный преобразователь, накопитель или логгер).

Именно поэтому в данной статье рассматриваются, в первую очередь, наиболее важные и принципиальные различия между автоматическими и традиционными способами контроля качества и обработки данных, определяющие специфику автоматизированного прогнозирования стока и других гидрологических процессов и явлений. Основной акцент сделан на описании ключевых принципов автоматической обработки данных с учетом рекомендаций Всемирной Метеорологической Организации (ВМО) и международных стандартов[1-6]; описание конкретных математических процедур вынесено за рамки данной статьи и будет представлено в последующих публикациях по автоматизированному прогнозированию стока.

В современных АСП данные собираются при помощи разнообразных автоматизированных средств сбора и передачи информации. Затем они должны пройти автоматический контроль качества, в результате которого эти данные могут быть автоматически откорректированы. В этом случае все внесенные изменения должны быть запротоколированы. Важно подчеркнуть, что обеспечению гарантии качества прогнозирования способствует внедрение наиболее передовых и прошедших достаточную апробацию методов проверки поступающих данных. Для этой цели ВМО рекомендует [4] национальным гидрологическим службам (при наличии необходимых ресурсов) рассмотреть возможность принятия программы управления качеством в виде, описанной, например, в [1].

Согласно [4], при традиционном сборе, обработке и использовании данных должны применяться подходы, прошедшие достаточную апробацию в оперативной деятельности прогностических подразделений национальных гидрометслужб. Гидролог обязан быть консервативным при проведении корректировки данных и выносить решение о необходимости изменений или добавлений только на основе строгих формальных критериев, основанных, скорее, на доказательствах, чем на предположениях. Если сделано то или иное предположение, оно остается в рамках ответственности пользователя, поскольку вся необходимая для этого информация может быть у него под рукой (например, в виде примечаний в полевом журнале или комментариев, хранящихся отдельно от базы данных). При автоматической обработке данных места

для субъективных предположений нет вообще: какое бы то ни было вмешательство в массив поступивших данных допустимо только на основе формальных критериев, установленных разработчиком АСП. Каждое изменение в данных должно быть записано таким образом, чтобы при необходимости можно было бы понять, что именно было сделано и почему. Для этой цели необходимо и достаточно иметь автоматически заполняемый журнал задокументированных процедур, с помощью которого можно будет проследить и проверить процесс работы с данными. Это условие тоже является требованием системы контроля качества разрабатываемых АСП.

Система контроля качества может включать те или иные механизмы (математические процедуры) корректировки входящих данных, выбор которых всецело зависит от научных пристрастий разработчика АСП. Однако они, тем не менее, должны соответствовать международным стандартам, подробно описанными в «Руководстве по климатологическим методам» [5].При разработке стандартов данных, целесообразно учитывать разрешение (дискретность, цену деления и т.п.) измерительных приборов, точность измерения физической величины, погрешность результата измерения и интегральную неопределенность измерительной процедуры [6]. Важно подчеркнуть, что уровень допустимой неопределенности обычно рассматривается в первую очередь. Как только он установлен, должны быть рассмотрены уровни неопределенности используемых методов, технологий и инструментов. Кроме того, в качестве формального критерия качества данных целесообразно использовать характеристики их полноты, от которой зависит полезность данных. Так, например, при разработке автоматической процедуры контроля качества необходимо определить разумные параметры ожидаемой полноты данных в виде, например, допустимого процентного соотношения пропущенных наблюдений. Это принцип может и должен быть положен в основу оптимизации автоматической наблюдательной сети, что позволит избежать необходимости заполнять пропуски предполагаемыми (в частности, интерполированными) значениями.

Первичная обработка данных

Первичная обработка заключается в подготовке данных к хранению в автоматической программно-реализованной базе данных (БД), в котором они будут доступны для использования в течение среднего или долгого срока. В зависимости от типа данных, первичная обработка может включать обработку графиков (например, графиков хода уровней воды, полученных с электронных самописцев) или обычную передачу файлов с незначительным редактированием или вообще без редактирования, в том виде, в котором эти данные получены с регистрирующего устройства (преобразователя или накопителя данных).Данные могут быть использованы и до такой обработки (например, данные о дистанционно измеренных уровнях воды), однако, конечный пользователь прогностической информации должен быть предупрежден, что данные не проверены и могут содержать ошибки. Рассмотрим основные элементы первичной автоматической обработки данных [4], которая должна выполняться при формировании непрерывных входных массивов прогностических моделей, используемых в АСП.

1. Предварительная проверка данных. Разница между предварительной проверкой и обнаружением ошибки довольно произвольная. Процедуры, которые в одной стра-

не считаются предварительной проверкой, в другой могут считаться обнаружением ошибки. Кроме того, степень использования компьютера при обработке данных может изменить понятие предварительной проверки. Для данных, собранных вручную, а затем переведенных в компьютерные файлы, термин «предварительная проверка» используется для обозначения процедур, предшествующих переводу данных в форму, пригодную для машинного считывания. Для данных, непосредственно собираемых в цифровой форме, требуется лишь незначительная проверка, которая отличается от обычного сравнения полученного массива данных с записями, выполненными вручную.Для телеметрических цифровых данных, предварительный контроль может быть незначительным или вообще не требоваться перед тем, как эти данные будут переданы пользователю или размещены в ассоциированной с АСП базой данных. В таких ситуациях пользователь должен быть предупрежден о том, что предоставленные данные являются непроверенными, и их следует использовать соответствующим образом. Даже в тех случаях, когда используется автоматическая процедура проверки данных, выполняется проверка лишь некоторых свойств данных (например, размаха, выбросов, шага, отсутствие некорректных символов или недостающих значений), поэтому пользователь должен знать ограничения и возможности применяемого способа предварительной проверки. Обычно телеметрические данные заносятся в файлы на станции или в другом безопасном месте системы сбора и обработки данных. Они могут быть отредактированы и обрести статус архивных или проверенных данных только после прохождения полного набора предварительных проверок (как это было описано ранее для регистрирующего устройства), процедуры обнаружения ошибок и тестирования.

2. Прослеживаемость и обработка. Система автоматической обработки данных в АСПдолжна обеспечивать:
   - Регистрацию данных после их сбора с целью подтверждения их существования и отслеживания их обработки;
   - Хранение резервных копий данных в их оригинальной форме;
   - Непосредственная идентификация отдельных массивов данных на различных стадиях их обработки;
   - Сравнение текущего статуса данных с изначальным, и их тестирование с точки зрения пригодности для использования;
   - Представление и хранение сведений о любых изменений данных;
   - Представление данных для проверки и контроля компетентными лицами, которые не имеют прямого отношения к процессу обработки данных.
3. Запись данных и отслеживание изменений в них. Данные, поступающие в АСП, должны быть зарегистрированы в автоматически заполняемом журнале, содержащем папки, которые систематизированы по типам станций или датчиков и представляют информацию в хронологическом порядке.
4. Идентификация и сохранение оригинальных записей. Все данные должны быть идентифицированы с использованием номеров станций, постов или датчиков и другими обязательными кодами. Для создания и хранения электронных данных необходимо иметь соответствующую систему наименования файлов и архив неизмененных оригинальных файлов данных. Рекомендуется видоизменить

имеющиеся файлы в надежный читаемый компьютером формат, который будет включать номера станций или датчиков и другие идентификаторы, чтобы в будущем не будет зависеть от программного обеспечении или средств, которые к тому времени устареют. Разработчикам АСП и БД (как являющихся частью АСП, так и автономных), рекомендуется обращать внимание на эту проблему при разработке и обновлении их файловых систем.

5. Преобразование данных с учетом автоматически идентифицированных ошибок. Такие ошибки могут быть обнаружены в результате сопоставления массива данных с массивами наблюдений на соседних автоматических станциях или постах или с данными наблюдений, полученными с соседних датчиков. Для корректировки таких ошибок БД должна автоматически выполнить необходимую корректировку на основе корреляции с данными наблюдений в соседних пунктах,или, в крайнем случае, при помощи линейной или более сложной интерполяции между зафиксированными значениями. Метод выполнения каждой корректировки должен быть отражен при помощи специальных кодов, используемых для протоколирования и документирования.
6. Накопление и интерполяция данных. Автоматические измерения многих переменных, в связи с их природной динамикой, должны проводиться в течение сравнительно короткого период, даже если в будущем они будут использоваться только в виде усредненных или суммарных значений за довольно длительные периоды времени. Современные базы временных рядов имеют все возможности для эффективного получения, обработки ихранения данных с любым уровнем агрегации (например, в виде средних величин за час, сутки, месяц и год). Частоту измерений, выполняемых при помощи автоматических датчиков, следует устанавливать с учетом стохастических характеристик наблюдаемого процесса, так, чтобы при наличии небольших пропусков данные можно было бы восстановить путем интерполяции. Каждый факт пропуска данных целесообразно доводить до оператора АСП или лица, принимающего решения о проверке функционирования автоматических датчиков.
7. Вычисление производных переменных. Производные переменные это те параметры, которые не измеряются непосредственно, а вычисляются с использованием других измерений (например, расход воды и испаряемость). Очевидно, что для данных, которые могут быть получены из основного фонда, не следует оставлять места для хранения в БД, поскольку оно ограничено.
8. Статус данных. Статус данных – это, как правило, автоматически определяемый код, обозначающий либо необходимость дополнительной обработки данных, либо отражающий полную готовность данных для дальнейшего использования в АСП. Например, автоматизированная система обработки данных Геологической службы США «ADAPS» имеет три уровня статуса: «рабочий», «находящийся на рассмотрении» и «одобренный». Системы БД, не имеющие таких опций, нуждаются во введении прав доступа к различным рабочим и архивным разделам, и другие привилегии, такие, например, как защита записей в файлах. Рекомендуется предоставлять право управления такой системой для каждой базы данных лишь одному человеку– например, системному администратору или специалисту, ответственному за контроль качества[4].

Вторичная обработка данных

Вторичная обработка представляет собой комплекс процедур, необходимых для производства данных в конвертированной или сокращенной форме (например, определение суточного количества осадков для отдельного водосбора, полученное из пиксельных радарных данных об отдельных дождях). Кроме того, термин «вторичная обработка» покрывает вторичное редактирование, выполняемое после более сложной проверки, а также заполнение пробелов в записях. При вторичной обработке данных должны использоваться следующие принципы [4, 7]:

– Изменения не должны производиться, если предположения не обоснованы с научной точки зрения;
– Произведенные изменения должны сопровождаться комментариями, которые могут быть автоматически записаны в специальный файл внесенных исправлений в БД;
– Общее правило автоматического заполнения пробелов – не заполнять пропуски искусственными данными и не получать пропущенные значения путем интерполяции. Любые приблизительные данные должны сопровождаться ссылками на соответствующие комментарии в БД. Исключения из общего правила не использовать синтетические данные или данные, полученные методом интерполяции, приводятся ниже. Например, пробелы в записях данных об уровнях воды могут быть заполнены методом линейной или нелинейной интерполяции, если выполнены все последующие условия:
    1) Речной сток находится в состоянии равномерного естественного спада или отличается неизменным уровнем воды;
    2) Установлено, что в период времени, соответствующий пропуску в данных, на водосбор не выпадало существенного количества осадков;
    3) Известно, что естественный гидрологический режим рассматриваемого водосборане искажается никакими внешними факторами (например, электростанциями или ирригационными гидротехническими сооружениями);
    4) Имеется непрерывность данных на каждой стороне от пропуска;
    5) В некоторых ситуациях, соседняя станция или соседний датчик может измерять те же или почти те же данные. В первом случае запись может быть заполнена так, как если бы это была резервная запись. Во втором данные могут быть введены, если погрешность меньше стандартного отклонения или если корреляция между наблюдениями на этих станциях (а также соотношение амплитуд) составляют не менее 0,99. Комментарий, содержащий подробности этого соотношения должен быть записан в соответствующий файл в БД;
    6) Рассматриваемая станция не расположена на озере, на уровенный режим которого влияют сейши или ветровые сгонно-нагонные явления.
– Заполнение пробела в оригинальных данных искусственными данными, полученными при помощи корреляции, недопустимо. Они могут, однако, быть использованы в тех случаях, когда пользователь знает о существовании погрешности, а также в случаях, когда восстановленные данные используются в качестве «входа» прогностической модели, требующей непрерывности «входа», «выход»

которой подлежит дальнейшей постобработке.В таких случаях рекомендуется помечать значения,прогнозируемые при помощи АСП и основанные на различных данных (фактических, обработанных, восстановленных и т.д.) соответствующими символами или элементами цветовой палитры;

– Пробел в данных по осадкам может быть интерполирован только тогда, если может быть установлено, что осадков не было в течение этого периода, за счет корреляции с другими осадкомерами внутри и вне водосбора, для которых установлена корреляция с коэффициентом не ниже 0.99.

Если необходимо заполнить пропуски, время, потраченное на приблизительное оценивание ряда перед обработкой, может принести большие дивиденды после анализа и обработки конечной информации. Обычно предпринимается попытка пополнить недостающие данные путем перекрестной корреляции с близлежащими станциями, особенно, теми, которые находятся в той же речной системе. При отсутствии такой возможности, можно использовать модели типа «осадки—сток», включая использование концептуальных моделей водосбора (например, моделей «Сакраменто» и «MLCM»). Все рассчитанные данные следует соответствующим образом обозначать «флажками» и хранить в отдельном архиве БД, а прогнозы стока, выпущенные на их основе, должны быть помечены соответствующим образом. Это позволит пользователю АСП самостоятельно оценивать степень надежности выпускаемых прогнозов стока.

Валидация и контроль качества

Процедуры валидации данных в общем случае заключаются в сравнении значений теста с исходными данными и часто присутствуют на нескольких уровнях первичной обработки, проверки данных и контроля качества. Несмотря на то, что методы компьютерного тестирования данных становятся более эффективными и мощными, следует понимать, что эта процедура никогда не будет автоматизирована до такой степени, что гидрологу не нужно будет проверять значения, помеченные флажками. В самом деле, для достижения лучшего результата пользователю АСП, возможно, придется постоянно изменять пороговые значения в программе. Кроме того, ему нужно будет регулярно принимать компетентное и взвешенное решение о том, принимать, отклонять или исправлять значения данных, отмеченные программой «флажками». Самые экстремальные значения могут оказаться правильными и очень важными при использовании гидрологических данных.

Не вызывает сомнений, что визуальная проверка графиков временных рядов опытным персоналом является быстрым и эффективным способом обнаружения аномальных значений. Поэтому большинство систем уточнения данных включают средства для построения таких графиков и выводят их на экран компьютеров, принтеров или плоттеров. Сопоставление данных с близлежащих станций это очень простой и эффективный путь мониторинга согласованности данных, полученных на этих станциях. Однако тщательная экспертная валидация в АСП возможна далеко не всегда, поэтому в данной работе будут рассмотрены лишь методики автоматической валидации.

Для того чтобы рассмотреть весь спектр методов, разработанных для систем автоматической валидации, полезно обратить внимание на абсолютные, относительные и физико-статистические виды проверки данных[4].

Абсолютная проверка означает, что данные или кодовые значения имеют такой диапазон изменений, вероятность превышения которого равна нулю. Так, например, географические координаты станции должны находиться в пределах страны, числа месяца могут изменяться только от 1 до 31 и т.д. Данные, не прошедшие этот тест, являются неверными. Обнаружить и исправить такие ошибки обычно очень просто.

Относительные проверки включают:

а) ожидаемый диапазон изменения переменных;

б) максимальную ожидаемую величину между двумя последовательными измерениями;

в) максимальную ожидаемую величину между значениями переменной на соседних станциях или датчиках.

На ранних стадиях развития и использования методик допустимые пределы ошибок рекомендуется сделать довольно широкими. Однако они не должны быть настолько широкими, чтобы это приводило к обнаружению трудно поддающегося обработке количества несогласованных значений. Эти пределы могут быть уменьшены по мере получения сведений относительно по вариации индивидуальных переменных. Поскольку эта процедура требует проведения фонового анализа исторических данных, ожидаемые диапазоны, необходимые для выполнения относительной проверки, должны быть рассчитаны для нескольких временных периодов, включая период проведения наблюдений. Это необходимо из-за того, что при значительном увеличении временного ряда уменьшается его дисперсия. В первую очередь, сравнивают ежедневные уровни воды с ожидаемым диапазоном суточных величин за текущий период времени, например за текущий месяц. Но, поскольку имеется вероятность того, что весь ряд значений был существенно (ошибочно) завышен или занижен, последующую проверку диапазона изменений нужно проводить за более продолжительный период времени. Таким образом, в конце каждого месяца следует сравнивать текущие среднемесячные значения со средним многолетним значением за этот месяц. Таким же образом в конце гидрологического года текущее среднегодовое значение сравнивается со средним многолетним. Этот способ применим ко всем временным гидрологическим рядам [4].

Метод сравнения каждого значения с предшествующим (метод (б)) наиболее применим к переменным, имеющим существенную внутри рядную корреляцию, например, к большинству рядов наблюдений за уровнем воды. В случае очень сильной внутри рядной корреляции (что характерно, например, для уровней подземных вод) можно выполнять сравнения за несколько периодов, как описано выше для метода (а). Ежесуточные данные наблюдений подземных вод можно, прежде всего, сравнить с ожидаемыми изменениями за день, а общее месячное изменение — с ожидаемым месячным.

Метод (в) представляет собой производную от метода (б), но он использует критерии ожидаемых изменений скорее в пространстве, чем во времени. Этот вид проверки наиболее эффективен для значений уровня (и расхода) воды в водотоках из одного и того же водосбора, хотя для сравнения данных о водотоках крупных бассейнов нужны специальные средства для накопления данных наблюдений, полученных с разных станций.

Для других гидрологических переменных полезность этого метода зависит от плотности наблюдательной сети, относительно ее пространственной изменчивости. Примером является преобразование суммарного количества осадков в безразмерные единицы при помощи отношения наблюденных величин к некоторому среднему многолетнему значению. Это приводит к уменьшению различий, вызванных характеристиками станций или используемых датчиков.

В ряде случаев целесообразно выполнение физико-статистической проверки, которая заключается в использовании регрессионных зависимостей между связанными переменными для прогнозирования ожидаемых значений. Примером такой проверки может служить сравнение уровня воды с суммарным количеством осадков или сравнение величины испарения, полученной с помощью испарителя, с температурой. Такая проверка обычно выполняется с данными, полученными со станций, расположенных в районах с редкой сетью, когда единственным средством проверки является сравнение со значениями связанных переменных, имеющих более плотную сеть. Другая категория физико-статистической проверки используется для подтверждения согласованности данных с общими физическими и химическими законами. Этот вид проверки широко применяется для данных о качестве воды. Большинство рассмотренных относительных и физико-статистических проверок основаны на использовании временных рядов, корреляции, множественной регрессии и методик обработки поверхностных данных.

Другой метод, который может быть использован для автоматической проверки относительных отклонений наблюдаемого элемента в течение какого-то периода, заключается в использовании различных типов математических взаимосвязей (например, полиномов). Рассчитанная величина сравнивается с наблюденной, и, если разница между ними не превышает предварительно установленного допустимого отклонения, данные считаются правильными, а если превышает, то этим данным требуется дальнейшая проверка.

Вопрос погрешности достаточно подробно рассмотрен в нескольких стандартах ISO, имеющих отношение к гидрометрическим методам. Публикация ISO «Руководство по выражению неопределенности измерений» [6]рекомендуется в качестве основного справочника по данной теме. Подробное руководство по оцениванию погрешности измерений расхода воды, важнейшего элемента контроля входных и выходных данных в гидрологических АСП, также представлено в [9, 10].

Работа выполнена в рамках мероприятия 1.2.2 Федеральной целевой программы «Научные и научно-педагогические кадры инновационной России» на 2009-2013 годы (государственный контракт № П1103 от 02 июня 2010 г.) по направлению «География и гидрология суши».

Литература

1. International Organization for Standardization, 2000: Quality Management Systems: Requirements. ISO 9001, Geneva.
2. International Organization for Standardization, 2005: Quality Management Systems: Fundamentals and Vocabulary. ISO 9000, Geneva.
3. International Organization for Standardization and International Electrotechnical Commission, 1995: Guide to the Expression of Uncertainty in Measurement. ISO/IEC Guide 98, Geneva.

4. WMO Guide 168
5. World Meteorological Organization, 1983: Guide to Climatological Practices. Second edition, WMO-No. 100, Geneva (http://www.wmo.int/pages/prog/wcp/ccl/guide/guide_climat_practices.html).
6. GuidetotheExpressionofUncertaintyinMeasurement, ISO, 1995
7. NationalInstituteofWaterandAtmosphericResearch, 1999
8. «Техническихправилах» ВМО №49 (TechnicalRegulations, WMO-No. 49), том III, Приложение, часть VIII.
9. World Meteorological Organization, 2006: Technical Regulations. Volume III – Hydrology, WMO-No. 49, Geneva.
10. World Meteorological Organization and Food and Agriculture Organization of the United Nations, 1985: Guidelines for Computerized Data Processing in Operational Hydrology and Land and Water Management. WMO-No. 634, Geneva.