УДК 004.9

М.Ю. Катаев, А.К. Лукьянов, А.А. Бекеров

Программная система накопления и манипулирования пространственно-временными данными

Рассмотрен принцип организации пространственно-временных структур разнородных данных (об атмосфере, поверхности Земли и др.). Описаны особенности работы с многомерными структурами данных. Представлен анализ требований к программной системе для работы с многомерными данными. Показана реализация программной системы и обсуждаются моменты для развития системы.

Ключевые слова: пространственно-временные структуры данных, доступ к данным, мониторинг, алгоритмы.

При изучении объектов реального объекта или явлений на первом этапе всегда происходит построение теоретической или эмпирической модели (как правило, в виде многопараметрической нелинейной динамической функции), которая близка к реальности (прямая задача). На втором этапе возникает задача определения параметров по значениям функции (обратная задача). Объекты мира являются сложносвязанными многопараметрическими функциями, и точность выделения объекта и определения его параметров зависят от знания информации о состояниях объекта во всех его проявлениях. Поэтому тем более точно будут определяться параметры изучаемого объекта, чем больше информации мы будем иметь об объекте и его изменениях в зависимости от тех или иных условий (иначе – более точно восстанавливать значения функции).

Задачи космического мониторинга земной поверхности и атмосферы [1, 2], где требуется знание пространственно-временной структуры параметров, как раз и относится к такого типа задачам, где наборы данных являются сложносвязанными и многомерными. Система «Земля+атмосфера» может быть описана многомерными наборами данных (поля влажности и других газов, аэрозоля, ветра, осадков и др.). Данные, получаемые с борта космических аппаратов, также являются многомерными. Все эти наборы данных можно структурировать по размерности: 1D (точка – значение), 2D (поверхность, линия), 3D (объем), 4D (многомерная поверхность). Поэтому обеспечение работы алгоритмов из такого набора данных является непростой с вычислительной точки зрения задачей [3, 4]. Особенностью таких задач является их строгая привязка к пространству $\{x$ – широта, y – долгота, z – высота$\}$ и/или ко времени $\{t\}$. Помимо этих основных параметров функции существует множество других сопутствующих параметров, сопровождающих данные наблюдений, которые уточняют основные характеристики функции, например, помимо времени, географической широты и долготы, необходимо знать угловое положение Солнца и космического аппарата относительно точки наблюдения, температуру и давление и др.

Эффективность использования пространственно-временных структур данных связана с решением конкретной задачи. В отдельных случаях не важна скорость работы алгоритмов и более существенна точность выборки (с учетом интерполяции) соответствующих массивов данных, в другом случае существенное значение имеет скорость выборки данных или минимизация требуемой памяти для хранения [5]. Все эти варианты требуют использования разного типа алгоритмов и соответственно различных наборов данных. Управление набором таких данных является сложной комплексной задачей, поэтому разработка программной системы работы с наборами данных для обеспечения информацией методов обработки спутниковых измерений является важной и актуальной для решения задач науки, сельского хозяйства и промышленности.

В данной работе представлен один из вариантов подготовки наборов данных, необходимых при решении прямых и обратных задач оптики атмосферы. Следует из некоторого множества разнородных пространственно-временных баз данных, которые являются глобальными (в пределах территории Земли) и представленными за некоторый определенный промежуток времени (в течение каждого дня за несколько лет), подготовить небольшие наборы данных для указанного времени и точки в пространстве.

Целью данной работы является описание разработанных в авторском коллективе подходов к накоплению и манипуляции большими, многомерными массивами данных при решении задач космического мониторинга в разрабатываемой программной системе.

Классификация пространственно-временных структур данных. Для решения задач космического мониторинга, расчета прохождения излучения и транспортных задач переноса вещества в атмосфере применяются разнообразные массивы данных. Многие из этих наборов разрабатывались для решения конкретных задач, и применение их в других задачах, как правило, затруднительно. Массивы данных специфически структурированы, сжаты в различных форматах данных (бинарный, HDF (Hierarchical Data Format), NetCDF (Network Common Data Form) и др.), имеют различные размерности (от 1D до 4D) и др. Для повышения эффективности работы систем с такими структурами данных необходимо учитывать не только особенность решаемой задачи, но и тип, размерность данных.

Некоторые научные организации создают на основе многолетних наблюдений, выполненных в космосе и на поверхности Земли, открытые для доступа данные (например, NASA, NCEP [6], ECMWF и др.). К таким наборам данных относятся: 4D-поля влажности ($\{x, y, z, t\}$), скорости и направления ветра, рельеф, типы почв (3D) и поверхности (2D) и др. Определенное количество этих наборов данных имеют глобальное описание (в пределах всей поверхности Земли).

Для космических исследований интеграция различных наборов данных, которые представляют количественные процессы, происходящие на поверхности Земли и в атмосфере, приводит к более точному пониманию самих процессов при визуальном исследовании, детализации и уточнению при решении прямых или обратных задач. При этом важной проблемой выступает организация большой по объему и разнородной информации в удобном для расчетов и конечного пользователя виде. Для выбранного направления исследований максимальный по объему массив данных является пятимерным (рис. 1).

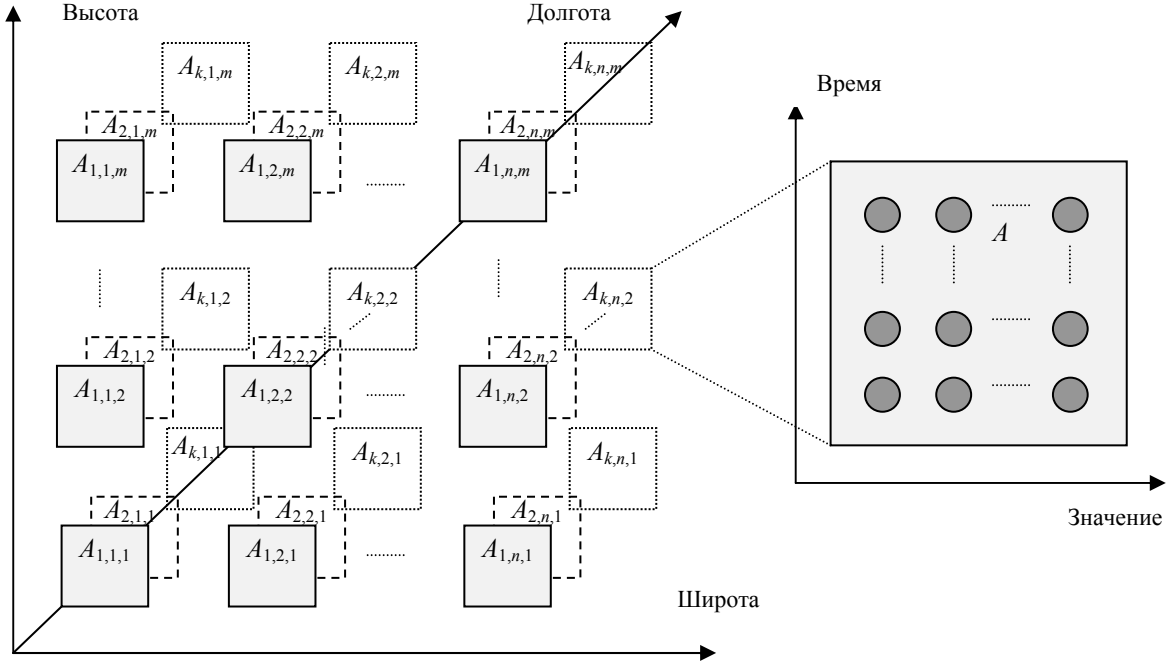

Рис. 1. Гиперкуб данных A космических измерений

Анализируя множество возможных вариантов выборки данных из структуры, представленной на рис. 1, можно получить типичные выборки, которые возможны для любых наборов данных:

1) точка (определение времени t и/или координат $\{x, y, z\}$);

2) линия (определение диапазона времени $t_2 - t_1$ и/или диапазона координат $\{x, y, z\}_2 - \{x, y, z\}_1$);

3) плоскость (определение времени t или диапазона времени $t_2 - t_1$ и/или диапазона координат углов прямоугольника $\{x, y, z\}_2$ и $\{x, y, z\}_1$) и другие варианты.

Из перечисленных выше вариантов выборки данных из структуры возникает большое количество подвариантов, связанных с конкретными заданными пользователем значениями атрибутов (например, для времени t_1 необходимо выделить на уровне поверхности Земли ($z = 0$) значения температуры поверхности (плоскость) в районе г. Томска (широта: 56°29' северной широты и долгота: 84°57' восточной долготы) размером 50×50 км). Эти запросы определяют вычислительную и алгоритмическую сложность ввиду наличия не единственной базы данных и разнотипности форматов, разных периодов времени и пространственных сеток (например, необходимо подготовить для за-

данного времени и географической точки наборы данных о температуре, давлении, скорости и направлении ветра, влажности и др.). Особенность запросов в том, что в базе данных информация хранится в одной физической размерности, а требуется получить информацию в другой, более удобной и понятной пользователю размерности.

Анализ требований к программной системе. Для направления исследований, связанного с космическими измерениями, каждая точка многомерного пространства является числом, а может представлять вектор значений (например, спектральную кривую или более сложную зависимость). Важнейшим требованием к подобной структуре данных является своевременная подготовка элементов этого пространства, необходимых для решения численных задач оптики атмосферы (прямые задачи) или восстановления параметров поверхности Земли или атмосферы (обратные задачи).

При работе с такими наборами данных возникает сложность, связанная с тем фактом, что при обработке больших объемов данных существует проблема поиска нужной точки пространства и представления результатов обработки (интерполяции) в виде соответствующих таблиц, определенного формата, подходящего для работы программ и анализа, визуального анализа, вывода. Процесс создания такого набора данных (выборки из больших таблиц) должен занимать время, которое не сказывается на основном вычислительном процессе, для которого данные и готовятся.

Анализ требований к программной системе, которая должна оперировать большими наборами данных, позиционировать в пространстве результаты и подключать программные единицы (библиотеки), показал, что она должна содержать следующие части программного обеспечения, необходимые для:

1) сбора информации;
2) сжатия и хранения;
3) доступа к накопленной информации;
4) организации информационного взаимодействия различных блоков программной системы;
5) визуализации пространственно-временных данных.

Наиболее удобным программным инструментом, в свете указанных выше характеристик программной системы, являются геоинформационные интернет-системы [7], построенные по принципу клиент-серверных технологий. Соответственно программная система имеет интерфейс доступа к функциям, обеспечивающим решение прикладных задач, а программные компоненты доступа к удаленным базам данных космической информации (например, расположенных на серверах NASA [https://podaac.jpl.nasa.gov/dataaccess]) средства администрирования самой системы и базы данных.

Одним из компонентов системы должна быть программная компонента для привязки данных – к системе географических координат. GeoServer [http://geoserver.org] является картографическим сервером с открытым исходным кодом, который среди многих прочих возможностей реализует следующие спецификации OGS: WMS, WFS, WCS. GeoServer реализует спецификацию WFS-T (WFS-Transaction). Это означает, что, используя GeoServer, можно не только получать данные для построения на их основе собственных карт, но также редактировать полученные данные с последующим автоматическим обновлением исходной информации на сервере. Среди поддерживаемых форматов значатся: JPEG, PNG, SVG, KML/KMZ, GML, PDF, ESRI Shapefile и др. Другой особенностью является поставляемая с GeoServer визуальная система управления файлами настроек и описания данных для программных проектов.

Разрабатываемая нами программная система реализована в виде веб-интерфейса и предоставляет пользователю возможность работать с многомерными данными, содержащими информацию о параметрах, описывающих Землю и атмосферу (рис. 2).

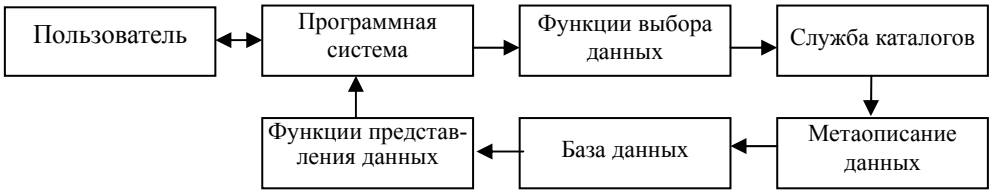

Рис. 2. Структура программной системы отображения многомерных данных

С общей информационно-технологической точки зрения программная система (ПС) рассматривается как многомодульная система, в составе которой выделяются следующие элементы:

1) функции выбора данных;

2) служба каталогов;
3) метаописание данных;
4) база данных (БД);
5) функции представления данных.

Рассмотрим подробнее каждый модуль отдельно. Функции выбора данных характеризует набор запросов пользователя, которые отражают тип и структуру выбираемых из БД данных. Служба каталогов формирует единое информационное пространство ПС, включающее: картографические слои, данные об атмосфере (температура, ветер и др.), данные о поверхности Земли (рельеф, типы поверхности, коэффициенты отражения и излучения и др.), документы (описание данных, методики и др.) и др. Организация структуры службы каталогов подразумевает логическое объединение разнородных данных физическими задачами (прямой или обратной) или просто желанием пользователя сделать выбор некоторых данных в заданном представлении (табличный, картографический или графический).

Программная система предоставляет пользователю универсальные средства обеспечения единообразного интерфейса ко всем наборам данных, автоматически настраивается на заданные функции выбора данных из базы данных и типовой набор стандартных функций работы с данными: поиск, добавление, изменение, удаление данных, интерполяция и др. Возможность решения этих задач в значительной степени обеспечивается использованием метаданных (структурированные данные, представляющие собой характеристики представленных в БД данных для целей их поиска и управления ими и др.).

База данных имеет простую логическую структуру, которая определяется типом и/или временем данных. Можно выделить, для примера, наборы данных о температуре атмосферы NCEP, которые формируются из файлов hdf, которые содержат информацию о профилях температуры на определенной пространственной сетке в течение года. Таким образом, структура хранения информации о температуре атмосферы связана с числом файлов, равным числу лет, за которые получены данные.

Функции представления данных связаны с запросами пользователя к ПС и могут быть связаны с задачей выборки, преобразования данных, пространственно-временного анализа и сравнения, решения задач расчета определенных характеристик и др. Все эти запросы преобразовываются в графический, картографический или табличный вид. Основной проблемой этого блока является интерполяция наборов данных для интервала значений, выбранных пользователем по заданным значениям сетки базы данных. Нами для этих целей применяются многомерные интерполяционные алгоритмы, разработанные в [8, 9].

Таким образом, для того чтобы пользователь получил данные в нужном ему виде, необходимо провести с базами данных ряд шагов: чтение, выделение участка с нужной информацией, интерполяция на сетку пользователя и запись в формате отображения данных. Чтение каждой базы данных требует своего отдельного модуля в силу того, что базы данных имеют различную размерность и способ хранения данных. После чтения и выделения необходимого пользователю участка данные становятся однообразными, представляя собой массивы размерности от одного до четырёх (рис. 3).

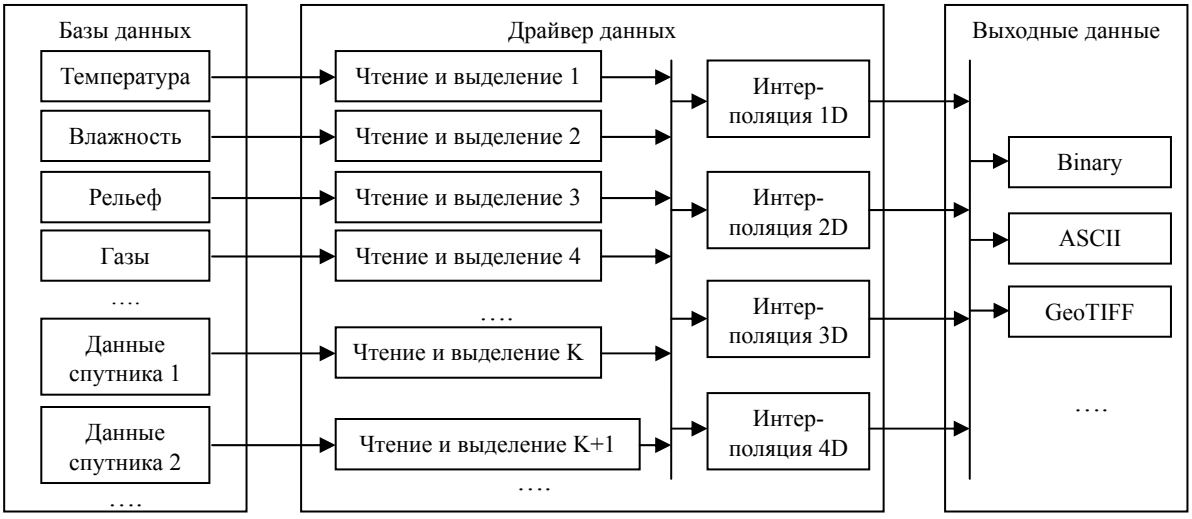

Рис. 3. Блок чтения и интерполяции многомерных данных

Формат выходных данных выбирается в зависимости от типа запроса и бывает бинарный (Binary), текстовый (ASCII) или в формате GeoTIFF (формат записи спутниковых изображений в формате TIFF, включая метаданные о географической привязке и характеристиках измерений). Файлы формируются с линейной и однотипной структурой данных. Это позволяет унифицировать передачу данных различным процедурам без преобразований.

Для примера, можно назвать несколько возможных вариантов выборки данных из набора данных, представляющих собой рельеф: 1) точка – значение высоты местности над уровнем моря, 2) линия – множество точек, показывающих изменение высоты местности, и 3) плоскость – изменение высоты местности на определенной территории. Во всех случаях, если искомая точка не попадает в узел сетки данных, возникает необходимость проводить интерполяцию. Поэтому в зависимости от типа случая необходимо выбирать соответствующее множество точек из набора данных для точного интерполирования. Представление выбранных и подготовленных данных в программной системе зависит от дальнейших операций: для отображения используется GeoTIFF, для продолжения преобразования, например расчета углов наклона поверхности, используется или текстовый (ASCII) или бинарный (Binary) вид.

Реализация программной системы. Нами для решения задач космического мониторинга территории Томской области разрабатывается интернет-ГИС программная система [10, 11]. Структура БД формируется из двух архивов: 1) данными, полученными с космических аппаратов, и априорными наборами данных и 2) информации о параметрах атмосферы и поверхности Земли за то же самое время, за которое получены данные в первой части. Задачей программной системы является экологический контроль территории (пожары, незаконные вырубки и здания, зона подтопления и др.). Внешний вид ПС приведен на рис. 4.

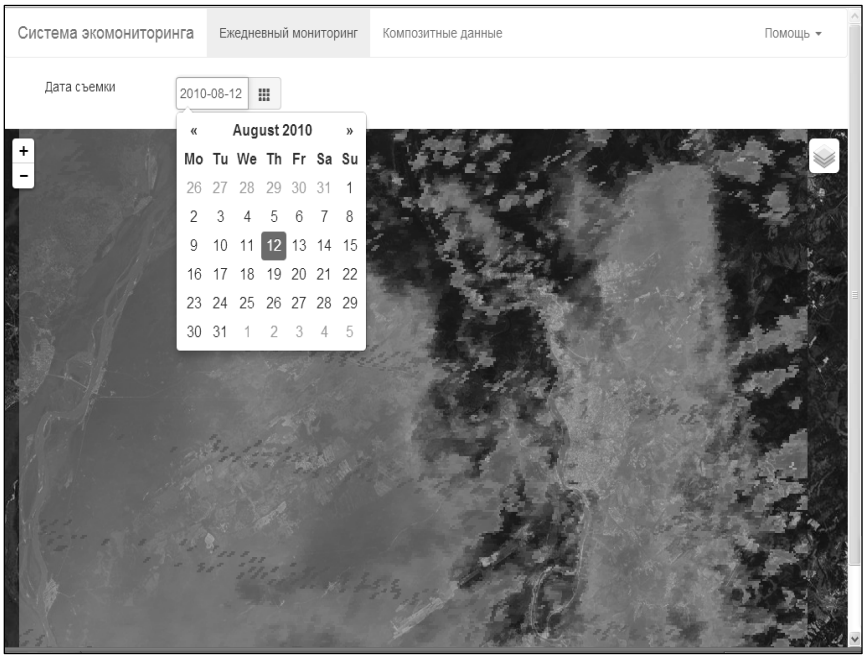

Рис. 4. Интерфейс программной системы (интернет-ГИС) для накопления, обработки и анализа спутниковых данных

В программной системе пользователь удален от наборов данных, которые необходимы для решения тех или иных задач, и выбирает только функции, реализующие запросы, на основе которых получаются результаты, представляемые в графическом или табличном виде. Для примера на рис. 4 представлен расчет вегетационного индекса NDVI для территории в районе г. Томска за определенную дату [11] на основе данных спутникового прибора MODIS [http://modis.gsfc.nasa.gov/]. Выбирая те или иные функции в ПС, пользователь может получать результаты и анализировать их, не задумываясь о сложности сбора, подготовки, преобразования, обработки и визуализации данных.

Заключение. Разработанная программная система для работы с пространственно-временными многомерными и разнородными наборами данных космического мониторинга позволяет решать задачи накопления, преобразования, манипулирования и визуализации. Найдены алгоритмические подходы к работе с многомерными и разнородными данными, заключающиеся в выборке и подготовке наборов линейных и однотипных по структуре данных, которые понимаются модулями программной системы, и другими программами. Выполнена разработка программной системы в виде интернет-ГИС для обработки данных спутникового прибора MODIS и получения наборов вегетационных индексов, которые связаны с климатическими параметрами (влажность, аэрозоль и др.), а также типами поверхности Земли (растительность, вода, почва и др.).

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №13-05-01036.

Литература

1. Кашкин В.Б. Дистанционное зондирование Земли из космоса. Цифровая обработка изображений / В.Б. Кашкин, А.И. Сухинин. – М.: Логос, 2001. – 322 с.

2. Чандра А.М. Дистанционное зондирование и географические информационные системы / А.М. Чандра, С.К. Гош. – М.: Техносфера, 2008. – 312 с.

3. Гулаков В.К. Пространственно-временные структуры данных / В.К. Гулаков, А.О. Трубаков, Е.О. Трубаков. – Брянск: БГТУ, 2013. – 215 с.

4. Гулаков В.К. Многомерные структуры данных / В.К. Гулаков, А.О. Трубаков. – Брянск: БГТУ, 2010. – 387 с.

5. Большаков А.А. Методы обработки многомерных данных и временных рядов / А.А. Большаков, Р.Н. Каримов. – М.: Горячая линия – Телеком, 2014. – 522 с.

6. NCEP/NCAR Reanalysis. Интернет-источник. – Доступ свободный: http://www.esrl.noaa.gov/psd/data/reanalysis/reanalysis.shtml

7. Samet R. Web based real-time meteorological data analysis and mapping information system // International Journal of Education and Information Technologies. – 2010. – T. 4. – Vol. 4. – P. 187–196.

8. Бахвалов Ю.Н. Метод многомерной интерполяции и аппроксимации и его приложения / Ю.Н. Бахвалов. – М.: Спутник+, 2007. – 108 с.

9. Зиновьев А.Ю. Визуализация многомерных данных / А.Ю. Зиновьев. – Красноярск: Изд-во Красноярского государственного технического университета, 2000. – 180 с.

10. Катаев М.Ю. Обнаружение экологических изменений природной среды по данным спутниковых измерений / М.Ю. Катаев, А.А. Бекеров // Оптика атмосферы и океана. – 2014. – Т. 27, № 7. – С. 652–656.

11. Катаев М.Ю. Геоинформационная система мониторинга экологического состояния территории по данным прибора MODIS / М.Ю. Катаев, А.А. Бекеров // Региональные проблемы дистанционного зондирования Земли. – Красноярск, 2014. – С. 120–123.

Катаев Михаил Юрьевич
Д-р техн. наук, профессор каф. автоматизированных систем управления (АСУ) ТУСУРа,
профессор Юргинского технологического института (филиала)
Национального исследовательского Томского политехнического университета
Тел.: +7-960-975-27-85, (382-2) 70-15-36
Эл. почта: kataev.m@sibmail.com

Лукьянов Андрей Кириллович
Ассистент каф. автоматизированных систем управления (АСУ) ТУСУРа
Тел.: +7-953-911-61-97
Эл. почта: hyena116@mail.ru

Бекеров Артур Александрович
Аспирант Института мониторинга климатических и экологических систем СО РАН (ИМКЭС)
Тел.: +7-985-852-57-65
Эл. почта: artur@bekerov.ru

Kataev M.Yu., Lukyanov A.K., Bekerov A.A.
Software system for storing and manipulating of the spatio-temporal data

A principle of organization of spatio-temporal structures of heterogeneous data about the atmosphere, Land surface, etc. is reviewed. Working with multidimensional data structures is described. An analysis of requirements for a software system to work with multidimensional data is presented. the implementation of software systems is shown and the opportunities for the development of the system are discussed.
Keywords: spatio-temporal data structures, data access, monitoring, algorithms.